View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Collection Of Biostatistics Research Archive

Collection of Biostatistics Research Archive

COBRA Preprint Series

Year 2009 *Paper* 54

Nonparametric Incidence Estimation From Prevalent Cohort Survival Data

Marco Carone* Masoud Asgharian†

Mei-Cheng Wang‡

*The Johns Hopkins University, mcarone@berkeley.edu

†McGill University, masoud@math.mcgill.ca

‡The Johns Hopkins University, mcwang@jhsph.edu

This working paper is hosted by The Berkeley Electronic Press (bepress) and may not be commercially reproduced without the permission of the copyright holder.

http://biostats.bepress.com/cobra/art54

Copyright ©2009 by the authors.

Nonparametric Incidence Estimation From Prevalent Cohort Survival Data

Marco Carone, Masoud Asgharian, and Mei-Cheng Wang

Abstract

Incidence is an important epidemiologic concept particularly useful in assessing an intervention, quantifying disease risk, and planning health resources. Incident cohort studies constitute the gold-standard in estimating disease incidence. However, due to material constraints, data are often collected from prevalent cohort studies whereby diseased individuals are recruited through a cross-sectional survey and followed forward in time. We discuss the identifiability of measures of incidence in the context of prevalent cohort survival studies and derive nonparametric maximum likelihood estimators and their asymptotic properties. The proposed methodology accounts for calendar-time and age-at-onset variation in disease incidence while also addressing common complications arising from the sampling scheme, hence providing flexible and robust estimates. We also discuss age-specific incidence and adjustments for temporal variations in survival. We apply our methodology to data from the Canadian Study of Health and Aging and provide insight into temporal trends in the incidence of dementia in the Canadian elderly population.

1 Introduction

In order to learn about survival with a disease or condition of interest, investigators often use prospective cohort studies, whereby groups of individuals are followed in time. One example is the incident cohort study, in which disease-free individuals are recruited and followed until death, loss to follow-up or study termination. Incident cohort studies provide bias-free survival data and allow for seemless inference regarding disease incidence in the target population. This type of study is considered to be the gold-standard in estimating incidence. However, the incident cohort study suffers from serious drawbacks. In order to accrue a sufficient number of disease cases, extensive enrollment and long follow-up periods are often required, particularly in relatively uncommon diseases. In practice, such requirements result in prohibitive logistic and economic costs to the investigators (see Szklo & Nieto (2000) and Rothman *et al.* (2008)).

As an alternative, investigators at times conduct prevalent cohort studies. A prevalent cohort study is conducted by recruiting prevalent individuals (living persons having experienced disease onset), determining their onset time (e.g. through medical records), and following them until death or potential censoring. By requiring minimal enrollment and reduced follow-up periods, the prevalent cohort study is more feasible to conduct. However, its design introduces systematic biases in the data-generation process, which, if ignored, can lead to grossly incorrect conclusions (e.g., as documented in Wolfson *et al.* (2001)). For example, individuals with longer disease durations are overrepresented in the sampling process; this bias in survival has been widely studied (see Cox & Oakes (1984), Tsai *et al.* (1987), Wang (1998) and Asgharian *et al.* (2002), for example).

If interest lies in disease incidence, it is not *a priori* clear whether the prevalent cohort

Hosted by The Berkeley Electronic Press

study provides the information needed to identify the underlying incidence process. At least two issues pose barriers to inference about incidence. The first regards the absence of a clearly-identified pool of disease-free individuals, susceptible to disease onset. In the incident cohort study, all individuals are initially disease-free; as time progresses however, some become diseased. Incidence can be gauged by studying the number for incident cases relative to the size of the susceptible pool. In the prevalent cohort study, the structure does not permit direct comparisons using available risk pools, as all individuals are diseased upon recruitment. The second difficulty is due to an inherent over-representation of individuals with onsets near the recruitment time: the earlier an individual's onset, the larger the hurdle (namely, surviving until recruitment time) this individual must overcome to be eligible for the study. This paper is concerned with nonparametric incidence estimation from prevalent cohort data. Interest in this problem arose from investigation into the Canadian Study of Health and Aging (CSHA), a study of dementia in the Canadian elderly population.

Variants of this problem have been discussed in the statistical and epidemiological literature in the past three decades. Early work by Miettinen (1976) and Freeman & Hutchison (1980), for example, established the celebrated epidemiological equations relating prevalence, incidence and mean duration under equilibrium conditions. Subsequent work by Alho (1992) generalized these results by considering exponentially-varying stable populations. Keiding (1991) provided a comprehensive look at incidence estimation from current-status data under Markovian structure; the absence of follow-up required the imposition of assumptions possibly difficult to verify in practice. Keiding *et al.* (1989) considered current-status data as well and suggested the use of inverse-weighting using background knowledge of historical mortality. Brookmeyer & Quinn (1995) focused on HIV infection incidence rate estimation

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art54

from cross-sectional data, and Alioum *et al.* (2005) also studied estimation of HIV infection incidence from prevalent cohort data using multistate Markov models. More recently, Addona *et al.* (2009) provided inferential results regarding incidence estimation under equilibrium conditions, whereby both population size and the number of diseased onsets per unit-time are assumed constant through time. We propose a flexible, nonparametric estimation framework allowing for calendar-time and age-at-onset variations in disease incidence and for temporal changes in population size, and accounting for common complications arising from the sampling scheme. Our framework is shown to provide optimal inferences under minimal assumptions. The methodology proposed is an example of a deconvolution problem, somewhat in the spirit of the *back-calculation method* elaborated in the setting of HIV infection (see Brookmeyer & Damiano (1989)).

The paper is organized as follows. The incidence measures to be studied are introduced in section 2. Identifiability of these measures and inference are discussed in sections 3 and 4. In section 5, the relaxation of certain assumptions made in sections 3 and 4 is outlined. In section 6, the methodology is used to decribe the incidence of dementia in the Canadian population from data collected as part of the Canadian Study of Health and Aging. Concluding remarks are provided in section 7.

COBRA
Collection of Biostatistics
Research Archive

2 Measures of incidence

The incidence of a disease refers to its occurrence in a susceptible population, and we distinguish between two related measures of disease incidence, each serving differing purposes. Let disease onset be a clearly-defined event and for simplicity, assume the disease of interest is irreversible. Suppose that in the target population onsets arise from a nonhomoge-

Hosted by The Berkeley Electronic Press

neous Poisson process $\{N(t), 0 < t < \infty\}$ (the onset point process) with intensity function $\lambda(t) = E[dN(t)|\mathcal{F}_{t^-}]/dt$ (where $\mathcal{F}_t$ is the natural filtration) and cumulative intensity function $\Lambda(t) = \int_0^t \lambda(u)du$. Here, the indexing variable t represents calendar-time and $t = 0$ is some (potentially arbitrary) time origin. In view of the memoryless property of the Poisson process, $\lambda(t)$ also equals the rate function $E[dN(t)]/dt$ of $\{N(t)\}$. This unconditional interpretation will be emphasized throughout.

The *incidence intensity*, an absolute instantaneous measure of disease occurrence, is the expected number of new disease cases per unit-time in a prespecified population. Probabilistically, this measure is the rate function $\lambda(t)$ of the onset point process. Its use is particularly important in public health policy, where resource planning and allocation require knowledge of the magnitude of disease burden over time. It suffers however from its dependence on the definition of a prespecified population, its scale being tied to the size of the underlying population. As such, it cannot be used directly to compare disease risk in different subpopulations. For example, a low-risk subpopulation may have the same incidence intensity as a high-risk subpopulation if, say, the former is much larger in size than the latter. Furthermore, secular variations in the incidence intensity may be due to changes in population size or in inherent disease risk.

For these reasons, the second measure considered, commonly referred to as the *incidence rate*, is a relative instantaneous measure of disease occurrence, obtained by standardizing the incidence intensity. Precisely, the incidence rate $r(t)$ is defined as

$$r(t) \;=\; \frac{\lambda(t)}{Q(t) - P(t)}\;, \tag{1}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

where $Q(t)$ and $P(t)$ are, respectively, the size of the target population and of its prevalent

http://biostats.bepress.com/cobra/art54

subpopulation at calendat-time t. It is important to note that every individual in the target population need not be at risk for the disease in question. Diseased individuals, although included in the size of the target population, may not contract the disease once again and consequently, should be discounted in the calculation of the at-risk population. This observation mandates standardization by $Q(t) - P(t)$ rather than simply $Q(t)$ as has been done in the literature (e.g., Keiding *et al.* (1989)). Of course, for rare diseases, this distinction is not critical; for many other diseases however, such as dementia in the elderly population, failure to perform this adjustment may lead to serious underestimation of the incidence rate. In view of this standardization, the incidence rate constitutes a true disease risk measure, as intended by epidemiologists and extensively used by public health scientists. It often serves as the fundamental metric on which etiologic and intervention-based research focuses. Concrete examples of its use include the identification of disease risk factors and the assessment of interventions aimed at disease prevention.

3 Incidence intensity: identifiability and inference

Suppose that survival with the disease of interest, denoted by X^0 (with survival function S and bounded support $(0, b)$), is independent of calendar-time of onset (see Section 5 where we relax this assumption). Suppose further that sampling occurs at a fixed timepoint $\tau > b$. Then, we may define the truncation variable, denoted by T^0 (with distribution function G), as the time elapsed between disease onset and recruitment. Individuals are deemed prevalent at recruitment if and only if their survival time X^0 exceeds their truncation time T^0. Denote by X and T the observable survival and truncation time random variables, respectively. Suppose that for prevalent individuals the residual lifetime $X - T$ is potentially

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

right-censored by a residual censoring variable C; denote the observed follow-up time and event indicator by $Y = \min(X, T + C)$ and $\Delta = \mathbb{I}_{Y=X}$, respectively. The introduction of censoring on the residual rather than full survival time reflects the fact that censoring is study-dependent and should thus act upon a subject's time under study alone. It is assumed that $(T^0, X^0 - T^0)$ is independent of D conditional upon $X^0 > T^0$. In view of biased sampling, only onsets associated with sufficiently long survival times may be observed: we define the w-observable onset point process $\{N_w^\star(t), 0 < t < \infty\}$ to be the counting process of onsets observable at calendar-time w, and note that $\{N_w^\star(t)\}$ is an independent $p_w(\cdot)$-thinning of $\{N(t)\}$ with thinning function $p_w(u) = S(w - u)$. It follows thus (see Schabenberger & Gotway (2005)) that $\{N_w^\star(t)\}$ is a nonhomogeneous Poisson process with intensity function $\lambda_w^\star(t) = S(w - t)\lambda(t)$ and cumulative intensity function

$$\Lambda_w^\star(t) = \int_0^t S(w - u) d\Lambda(u) \ . \tag{2}$$

3.1 Identifiability

The general problem considered here is that of determining sufficient conditions under which estimation of $\lambda(\cdot)$ (or equivalently, deconvolution of (2)) is possible using a sample from the process $\{N_\tau^\star(t)\}$. With only survival data of prevalent individuals sampled at recruitment, the intensity function $\lambda(t)$ (or equivalently, the centered cumulative intensity function $\Lambda_c(t)$ defined as $\Lambda(t) - \Lambda(\tau - b)$ for $\tau - b \leq t < \tau$ and 0 otherwise) is identified only up to a constant of proportionality on $(\tau - b, \tau)$. This fact follows from noting that $d\Lambda(t)$ is proportional to

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art54

$dG(\tau - t)$ (see Daley & Vere-Jones (2003) and Note in Appendix) and consequently, that

$$\Lambda_c(t) \;=\; \kappa\,(1 - G(\tau - t)) \tag{3}$$

for some constant $\kappa > 0$. Since identification of G, the truncation distribution, is possible on $(0, b)$ (see Wang (1991)), our claim follows. Of course, without further assumptions, it is impossible to nonparametrically identify $\Lambda(\cdot)$ on $(0, \tau - b)$ as individuals with onset earlier than $\tau - b$ are not subject to sampling.

Full nonparametric identifiability of Λ_c can be attained with an additional requirement on the data-collection process. If an estimate of population disease prevalence at recruitment is available, either through external sources or as provided by the sampling framework, the constant of proportionality κ may be identified. This fact follows from the combination of (2) and (3) into

$$\Lambda^\star_\tau(t) \;=\; \kappa \int_0^t S(u) dG(u)\ , \tag{4}$$

and the observation that $\Lambda^\star_\tau(\tau)$ is the mean population prevalence at recruitment. Thus, the centered cumulative intensity function may be written as

$$\Lambda_c(t) \;=\; \frac{\Lambda^\star_\tau(\tau)}{\int_0^\infty S(u) dG(u)}\,(1 - G(\tau - t))\ , \tag{5}$$

a functional of identifiable arguments.

In order to simplify matters, epidemiologists often assume the so-called *stable disease conditions*, which require stationarity of the onset point process and stability of the size of the target population. Together, these requirements imply both constancy of disease prevalence over time and uniformity of the truncation distribution. Under such conditions, equation (5)

Hosted by The Berkeley Electronic Press

reduces to

$$\text{prevalence} \;=\; \text{incidence} \;\times\; \text{mean duration}\;,$$

a celebrated result in epidemiology (see, for example, Albert *et al.* (1978a), Albert *et al.* (1978b) and Louis *et al.* (1978)). The framework provided in this paper may thus be seen as a generalization of this result, allowing for arbitrarily-varying incidence and population sizes over calendar-time.

In many prevalent cohort studies, sampling of the prevalent individuals is performed cross-sectionally, in the sense that a simple random sample of the population is selected and each subject in this sample is assessed for prevalent disease. Thus, in addition to providing a sample of prevalent individuals, each contributing (biased) onset and survival data, an estimate of disease prevalence in the population is automatically obtained. In view of the above discussion, identifiability of the intensity function on $(\tau - b, \tau)$ is assured. Considerable weakening of this assumption will be considered in section 5.

The centered cumulative intensity function lends itself to easy interpretation. Indeed, for $\tau - b \leq t_1 < t_2 < \tau$, the difference $\Lambda_c(t_2) - \Lambda_c(t_1)$, identically equal to $\Lambda(t_2) - \Lambda(t_1)$, is the mean number of disease onsets occurring in the target population between calendar-times t_1 and t_2. The centered cumulative intensity function $\Lambda_c(t_2)$ provides the same interpretation but fixing $t_1 = \tau - b$.

3.2 Estimation

From the discussion above, a natural plug-in estimator for $\Lambda_c(t)$ emerges as

$$\hat{\Lambda}_c(t) \;=\; \frac{\hat{\Lambda}^{\star}_{\tau}(\tau)}{\int_0^{\infty} \hat{S}(u) d\hat{G}(u)} \Big(1 - \hat{G}(\tau - t)\Big) \;. \tag{6}$$

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art54

In the above, the estimator $\hat{\Lambda}^{\star}_{\tau}(\tau)$ is simply the estimated population prevalence $n_d n_{pop}/n_s$, where n_s, n_d and n_{pop} are, respectively, the number of individuals sampled, the number of such individuals found diseased, and the size of the target population at recruitment. The estimators $\hat{S}$ and $\hat{G}$ are, respectively, the truncation product-limit estimator (Tsai *et al.* (1987)) and Wang's inverse-weighted estimator of the truncation distribution (Wang (1991)). More explicitly, we have that

$$\hat{S}(u) = \prod_{i=1}^{n} \left\{ 1 - \frac{dN(Y_i)}{R(Y_i)} \right\}^{N_i(u)\Delta_i} \quad \text{and} \quad \hat{G}(t) \;=\; \sum_{i=1}^{n} \frac{\mathbb{I}_{(-\infty,T_i]}(t)}{\hat{S}(T_i)} \;/\; \sum_{i=1}^{n} \frac{1}{\hat{S}(T_i)} \;,$$

where $N_i(u) = \mathbb{I}_{(-\infty,u]}(Y_i)$, $N(u) = \sum_{i=1}^{n} N_i(u)$ and $R(u) = \sum_{i=1}^{n} \mathbb{I}_{(T_i,X_i)}(u)$. Since each of these estimators are NPMLE for their respective targets, the invariance property of maximum likelihood estimation guarantees that the proposed estimator is the NPMLE of the centered cumulative intensity function.

Substitution of the explicit form of $\hat{\Lambda}^{\star}_{\tau}(\tau)$ and $\hat{G}$ into (6) provides a simple and intuitively-appealling form for $\hat{\Lambda}_c(t)$; specifically, (6) reduces to

$$\frac{n_{pop}}{n_s} \sum_{i=1}^{n_d} \frac{\mathbb{I}_{(\tau-t,\infty)}(T_i)}{\hat{S}(T_i)} \;, \tag{7}$$

where $T_1, T_2, ..., T_{n_d}$ are the observed truncation times.

If interest lies in estimating the intensity function itself (rather than its integrated form), the above estimator may be used in conjunction with smoothing-based differentiation techniques.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

3.3 Asymptotic properties

The estimator proposed for the centered cumulative intensity function above exhibits desirable large-sample behavior. Our first theorem establishes its consistency, while the second theorem provides its asymptotic law. The proofs of these theorems are provided in the Appendix.

Theorem 1. *Under the assumptions stated in Sections 2 and 3, the estimator* $\hat{\Lambda}_c(t)$ *is uniformly strongly consistent for the true underlying centered cumulative intensity function* $\Lambda_c(t)$ *over* $(\tau - b, \tau)$*, that is,*

$$\sup_{\tau - b < t < \tau} |\hat{\Lambda}_c(t) - \Lambda_c(t)| \xrightarrow{a.s.} 0 \ .$$

Denote by $\mathbb{H}_n(\cdot)$ the empirical process $\sqrt{n}(\hat{H}_n(\cdot) - H(\cdot))$ associated to an estimator $\hat{H}_n(\cdot)$ of $H(\cdot)$.

Theorem 2. *Under the assumptions of Theorem 1, the normalized cumulative intensity process* $\sqrt{n_s}(\hat{\Lambda}_c(t) - \Lambda_c(t))$ *converges weakly to a mean-zero Gaussian process with covariance function* Σ *given by*

$$\begin{aligned}
\Sigma(s,t) &= \frac{\Lambda^\star_\tau(\tau)(1 - G(\tau - s))(1 - G(\tau - t))}{\beta^2} \left\{ 1 + \frac{n_{pop}}{\beta^2}\nu - \Lambda^\star_\tau(\tau) \right\} + \\
&\quad \frac{\Lambda^\star_\tau(\tau) n_{pop}}{\beta^2} \left\{ \sigma^2_{\mathbb{G}_{n_d}}(\tau - t, \tau - s) + \frac{1 - G(\tau - s)}{\beta}\phi(t) + \frac{1 - G(\tau - t)}{\beta}\phi(s) \right\} \ , \\
\text{where} \quad \beta &= \int_0^\infty S(u) dG(u) \ , \\
\nu &= \int_0^\infty \int_0^\infty \sigma^2_{\mathbb{S}_{n_d}}(u,v) dG(u) dG(v) - 2 \int_0^\infty \int_0^\infty \sigma_{\mathbb{S}_{n_d}, \mathbb{G}_{n_d}}(u,v) dG(u) dS(v) + \\
&\quad \int_0^\infty \int_0^\infty \sigma^2_{\mathbb{G}_{n_d}}(u,v) dS(u) dS(v) \ , \\
\phi(w) &= \int_0^\infty \int_0^\infty \left\{ \sigma_{\mathbb{S}_{n_d}, \mathbb{G}_{n_d}}(u, \tau - w) - \sigma^2_{\mathbb{G}_{n_d}}(\tau - w, v) \right\} dG(u) dS(v) \ ,
\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art54

and $\sigma^2_{\mathbb{S}_{n_d}}(u,v)$, $\sigma^2_{\mathbb{G}_{n_d}}(u,v)$ *and* $\sigma_{\mathbb{S}_{n_d},\mathbb{G}_{n_d}}(u,v)$ *are, respectively, the asymptotic covariance between* $\mathbb{S}_{n_d}(u)$ *and* $\mathbb{S}_{n_d}(v)$*, between* $\mathbb{G}_{n_d}(u)$ *and* $\mathbb{G}_{n_d}(v)$*, and between* $\mathbb{S}_{n_d}(u)$ *and* $\mathbb{G}_{n_d}(v)$*, each provided in Wang (1991).*

In view of the above, approximate confidence intervals and bands may be obtained by estimating the above covariance function $\Sigma(s,t)$ via substitution by appropriate empirical estimates. However, given the covariance function's rather intricate form, use of properly-specified boostrap resampling is likely to be more expedient and accurate. Theorems 1 and 2 justify the use of the bootstrap.

In the provided setting, the bootstrap must be performed in two stages, similar to the data-generation process. First, the number of prevalent individuals sampled n_d^{boot} should be generated from a Binomial distribution with size n_s and success probability n_d/n_s. Second, the survival data of the n_d prevalent individuals should be resampled with replacement to obtain a sample of n_d^{boot} survival triplets. The bootstrap sample obtained is then comprised of the generated diseased sample size n_d^{boot} and the resampled survival data.

The asymptotic behavior of estimators of the intensity function falls into the realm of density estimation (see for example Ramlau-Hansen (1983)) and as such, is beyond the scope of this paper. We defer its exposition to future work.

4 Incidence rate: identifiability and inference

Using (1) and the observation that $\Lambda_t^\star(t)$ is the mean disease prevalence in the target population at calendar-time t, we may write the centered cumulative incidence rate (initiating at

COBRA
A BEPRESS REPOSITORY
Research Archive

Hosted by The Berkeley Electronic Press

$\tau - b$) as

$$R_c(t) \;=\; \int_0^t \frac{d\Lambda_c(u)}{Q(u) - \Lambda_u^\star(u)} \;. \tag{8}$$

It is interesting to note that (8) reduces to another celebrated epidemiological relationship, namely

$$\text{prevalence odds} \;=\; \text{incidence rate} \;\times\; \text{mean duration} \;,$$

under the stable disease conditions.

4.1 Identifiability

Unlike the intensity function, the incidence rate is not identifiable from prevalent cohort data alone. If the size of the prevalent population $\Lambda_t^\star(t)$ through time is available, nonparametric identifiability is guaranteed. External sources of information must be consulted to obtain estimates of $Q(t)$ and values for $\Lambda_t^\star(t)$ through time. The former may usually be obtained readily from census data, for example. The latter, however, is usually more problematic. If such information is not available, some ad-hoc methods should be devised to estimate it from the available data. Although strictly correct, it is slightly misleading to claim that $\Lambda_t^\star(t)$ is unidentifiable from prevalent cohort data. Indeed, depending on the value of t, some information about $\Lambda_t^\star(t)$ may be recovered from the data. For $\tau - b < t \leq \tau$, (2) may be decomposed as

$$\begin{aligned}\Lambda_t^\star(t) \;&=\; \int_0^t S(t-u)d\Lambda(u) \;=\; \int_{t-b}^t S(t-u)d\Lambda(u) \\ &= \int_{t-b}^{\tau-b} S(t-u)d\Lambda(u) + \int_{\tau-b}^t S(t-u)d\Lambda(u) \;.\end{aligned}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

The second integral is identifiable from the available data; the first is not due to its integrator.

http://biostats.bepress.com/cobra/art54

4.2 Estimation

If historical prevalence data are available, then estimation of the incidence rate over time is relatively straightforward. Indeed, the plug-in estimator for $R_c(t)$ is

$$\hat{R}_c(t) \;=\; \int_0^t \frac{d\hat{\Lambda}_c(u)}{Q(u) - \Lambda^\star_u(u)} \;, \tag{9}$$

which simplifies to

$$\frac{n_{pop}}{n_s} \sum_{i=1}^{n_d} \frac{\mathbb{I}_{(\tau-t,\infty)}(T_i)}{\hat{S}(T_i)(Q(\tau - T_i) - \Lambda^\star_{\tau-T_i}(\tau - T_i))} \;. \tag{10}$$

Without historical prevalence information however, some form of extrapolation is inevitable. One possible extrapolatory approach consists of specifying a model, say $\lambda(t;\theta)$, for the intensity function between $(\tau - 2b, \tau)$ and using some distance-minimization technique and the above NPMLE for $\Lambda_c(t)$ over $(\tau - b, \tau)$ to obtain parameter estimates $\hat{\theta}$. This model-based estimate of the intensity over $(\tau - 2b, \tau - b)$ combined with the nonparametric estimates of the intensity over $(\tau - b, \tau)$ and of the survival function could then be used to estimate $\Lambda^\star_t(t)$: specifically, we set

$$\hat{\Lambda}^\star_t(t) \;=\; \int_{t-b}^{\tau-b} \hat{S}(t-u)\lambda(u;\hat{\theta})du + \int_{\tau-b}^{t} \hat{S}(t-u)d\hat{\Lambda}_c(u) \;.$$

Because any extrapolatory method should be chosen and fine-tuned in consideration of the particular application considered, the aymptotic details of such procedures are omitted.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

4.3 Asymptotic properties

Under the availability of historical prevalence estimates, inference about the centered cumulative incidence rate builds upon our knowledge of the limiting behavior of the centered cumulative intensity function. We assume in the sequel that $Q(u) - \Lambda_u^\star(u) > 0$ for all $u \in [0, \tau]$. The following results describe the asymptotic behavior of $\hat{R}_c$.

Theorem 3. *Under the assumptions of Theorem 1 and strict positivity of* $Q(u) - \Lambda_u^\star(u)$ *on* $[0, \tau]$*, the estimator* $\hat{R}_c(t)$ *is uniformly strongly consistent for the true underlying centered cumulative intensity function* $R_c(t)$ *over* $(\tau - b, \tau)$*, that is,*

$$\sup_{\tau - b < t < \tau} |\hat{R}_c(t) - R_c(t)| \xrightarrow{a.s.} 0 \ .$$

Theorem 4. *Under the assumptions of Theorem 3, the cumulative incidence rate empirical process* $\sqrt{n_s}(\hat{R}_c(t) - R_c(t))$ *converges weakly to a mean-zero Gaussian process with covariance function* Ψ *given by*

$$\Psi(s,t) \ = \ \int_0^t \int_0^t \frac{\Sigma(u,v)}{(Q(u) - \Lambda_u^\star(u))(Q(v) - \Lambda_v^\star(v))} \, d\Lambda_c(u) d\Lambda_c(v) \ ,$$

where Σ *is the covariance function provided in Theorem 2.*

As noted earlier, estimation and inference regarding the incidence rate, as opposed to the centered cumulative incidence rate, may be dealt with using a variety of smoothing differentiation techniques.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

5 Extensions of the methodology

5.1 Covariates and age-specific incidence

The theory presented in sections 3 and 4 concerns one-sample estimation. In practice, it is usually of interest to determine not only population-averaged measures of incidence, but rather measures relevant to particular subgroups of the population. For covariates of only finitely many levels, the methodology proposed may be readily adapted.

Suppose Z is a covariate taking values in a finite set. Then, the centered cumulative intensity function pertaining to $Z = z$ may be written as

$$\Lambda_{c,z}(t) \;=\; \frac{\Lambda^{\star}_{\tau,z}(\tau)}{\int_0^{\infty} S_z(u) dG_z(u)} \left(1 - G_z(\tau - t)\right) \;, \tag{11}$$

where S_z and G_z are the stratum-specific survival function and truncation distribution function, respectively, and $\Lambda^{\star}_{\tau,z}(\tau)$ is the stratum-specific prevalence at time τ. Each of S_z and G_z may be estimated by considering the subset of the prevalent cohort data for which $Z = z$. Furthermore, at least two approaches may be used to estimate $\Lambda^{\star}_{\tau,z}(\tau)$: we may either take

$$\tilde{\Lambda}^{\star}_{\tau,z}(\tau) = n_{d,z} n_{pop,z} / n_{s,z} \quad \text{and} \quad \hat{\Lambda}^{\star}_{\tau,z}(\tau) = n_{d,z} n_{pop} / n_s \;,$$

where $n_{d,z}$, $n_{s,z}$ and $n_{pop,z}$ are, respectively, the number of stratum-specific prevalent inviduals sampled, the number of stratum-specific invidivuals sampled and the size of the stratum-specific subset of the population. Assuming that either of n_{pop} and $n_{pop,z}$ are available from external sources, independence between sampling and the covariate of interest ensures that both estimators are consistent for the stratum-specific prevalence of disease at time τ. Usual

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

subgroup analysis consists of conditioning on any particular covariate value, and it is clear that $\tilde{\Lambda}^{\star}_{\tau,z}(\tau)$ is the corresponding subgroup estimator of $\Lambda^{\star}_{\tau,z}(\tau)$. We argue however that $\hat{\Lambda}^{\star}_{\tau,z}(\tau)$ is more appropriate than $\tilde{\Lambda}^{\star}_{\tau,z}(\tau)$ in that it alone allows extension to age-specific incidence estimation. Indeed, age-at-onset, despite being an often crucial quantity, is not a well-defined covariate in that it does not partition the whole population: this observation is a consequence of age-at-onset being defined only in individuals having experienced disease onset. Thus, while $n_{s,z}$ and $n_{pop,z}$ are not defined for age-at-onset, all terms in $\hat{\Lambda}^{\star}_{\tau,z}(\tau)$ are. An estimator of $\Lambda_{c,z}(t)$ is thus

$$\hat{\Lambda}_{c,z}(t) \;=\; \frac{n_{pop}}{n_s} \sum_{i:Z_i=z} \frac{\mathbb{I}_{(\tau-t,\infty)}(T_i)}{\hat{S}_z(T_i)} \;. \tag{12}$$

It is important to not confuse $\Lambda^{\star}_{\tau,z}(\tau)$, the prevalence of age-specific disease, with age-specific prevalence. The first considers individuals with onsets in a particular age group, while the second concerns individuals with prevalent disease *during* this age group. This distinction should not be understated.

5.2 Stratified sampling

The theory above rests on the assumption that individuals were recruited via simple random sampling from the target population. This assumption allows estimation of population prevalence, for example, from sample prevalence. In many cases however, the study design incorporates stratification in the sampling scheme so as to maximize recruitment of cases. With appropriate knowledge of population characteristics, it is possible to recover correct population estimates from sample quantities through reweighting.

Suppose that Y is the stratifying covariate, defined for all individuals, either diseased

http://biostats.bepress.com/cobra/art54

at recruitment or not, and that its range is a finite set $\mathcal{S}$. Let Z be a covariate of interest and D indicate disease status. Then, the joint stratum-specific prevalence probability may be written as

$$\Pr(D=1, Z=z) \;=\; \sum_{y\in\mathcal{S}} \Pr(D=1, Z=z|Y=y)\Pr(Y=y)\;, \tag{13}$$

where $\Pr(D=1, Z=z|Y=y)$ is directly estimable from the data at hand and $\Pr(Y=y)$ should be obtained from external sources. Of course, if the stratifying variable is independent of both disease status and covariate Z, there is no need for adjustment. It follows from (13) that the adjusted stratum-specific population prevalence estimate is

$$\hat{\Lambda}^{\star}_{\tau}(\tau) \;=\; n_{pop} \sum_{y\in\mathcal{S}} \left[\frac{n_{d,z|y}}{n_{s|y}} \Pr(Y=y)\right]\;,$$

where $n_{d,z|y}$ and $n_{s|y}$ are, respectively, the number of stratum-specific diseased individuals and the number of sampled individuals, each specific to stratifying value y. Thus, for covariate Z, the correction factor for unadjusted prevalence estimates corresponding to $Z=z$ is

$$\xi_z \;=\; \sum_{y\in\mathcal{S}} \left[\frac{n_{d,z|y}}{n_{d,z}} \frac{n_s}{n_{s|y}} \Pr(Y=y)\right]\;.$$

This measure provides an indication of the underrepresentation of covariate level $Z=z$ imparted by the stratification in the sampling process.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

5.3 Temporal trends in survival

Despite incorporating assumptions much less stringent than usually encountered in the relevant literature, the theory presented above nonetheless builds upon a set of assumptions which may fail to hold in practice. One such assumption is the independence between survival and onset time, or rather, the absence of any temporal trend in survival. This assumption may however be relaxed substantially.

In view of recent work (Cheng *et al.* (2007)), it is impossible to fully identify the onset-conditional survival distribution nonparametrically without specification of the dependence structure between survival and onset time. And in most cases, we may argue that such a general estimation framework may be overly cumbersome for the considered objectives. Rather, using that onset times and truncation variables are in bijection given a fixed sampling time, we consider incorporating temporal trends in survival via a proportional hazards model relating survival and truncation (see Wang *et al.* (1993)).

We denote the hazard rate at time x associated to individuals with truncation time t by $h(x|t)$. Consider the one-parameter regression model of survival conditional upon truncation

$$h(x|t) \;=\; h_0(x) \exp\left\{\gamma \phi(t)\right\} \,, \tag{14}$$

for some specified univariate trend function ϕ and an unspecified baseline hazard rate h_0. The choice of ϕ should, in practice, be motivated by the scientific application of interest. For example, a disease for which survival is believed to have varied only insidiously over time might warrant $\phi(t) = t$ or $\phi(t) = \log t$, while a disease for which a marked change in survival occurred (e.g, due to an innovative treatment) could be modeled via some step

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art54

function $\phi(t) = \mathbb{I}_{[0,\tau_0)}(t)$ for a specified τ_0. Apart from adjusting for potential trends in survival, this model allows quantification of temporal trends along with testing of the no-trend hypothesis $\mathcal{H} : \gamma = 0$. Estimation and inference based on the above model can be seemlessly performed through usual risk-set methods to account for truncation. Under potential dependence between onset time and survival, equation (5) still holds replacing $S(u)$ by $S(u|u)$, where $S(x|t) = \Pr(X^0 > x|T^0 = t)$ is the survival function corresponding to individuals with truncation value t (or onset time $\tau - t$) evaluated at x. Then, we obtain the estimator

$$\hat{\Lambda}_c^{onset}(t) \;=\; \frac{\hat{\Lambda}_\tau^\star(\tau)}{\int_0^\infty \hat{S}_{\hat{\gamma}}(u|u)d\hat{G}(u)} \Big(1 - \hat{G}(\tau - t)\Big) \;, \tag{15}$$

where $\hat{S}_{\hat{\gamma}}(u|u)$ is the model-based survival estimator described above and $\hat{G}$ is some suitable estimator of the truncation distribution function. It is possible to extend the work of Wang (1991), which assumes independence between the failure time and truncation random variables, to account for the considered dependence by noting that

$$G(t) \;\propto\; \int_{u \le t} \frac{dG^{obs}(u)}{S(u|u)} \;,$$

with G^{obs} the conditional distribution function of T^0 given $T^0 \le X^0$ (estimable directly by the empirical cdf). We obtain an appropriate estimator for G to be

$$\hat{G}(t) \;=\; \sum_{i=1}^{n} \frac{\mathbb{I}_{(\tau-t,\infty)}(T_i)}{\hat{S}_{\hat{\gamma}}(T_i|T_i)} \;/\; \sum_{i=1}^{n} \frac{1}{\hat{S}_{\hat{\gamma}}(T_i|T_i)} \;. \tag{16}$$

This estimator can be shown to be consistent and asymptotically efficient under correct specification of the dependence structure between onset and failure times.

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

6 The Canadian Study of Health and Aging

In 1989, researchers from the University of Ottawa, in conjunction with Health Canada designed the Canadian Study of Health and Aging (CSHA), a nationwide multicenter longitudinal study aiming to describe the current epidemiology of dementia in Canada. The shift in population age distributions, the consequent change in the occurrence of geriatric diseases and its implied impact on health services utilization motivated the research efforts involved in the design of this study. Most pressingly, the researchers wished to determine the prevalence, incidence and key risk factors of various dementia, including Alzheimer's disease, in several subpopulations of Canada. See McDowell *et al.* (2005a) for more details.

The study design included three distinct stages, referred to as CSHA-1, CSHA-2 and CSHA-3, chronologically. The first stage of the study took place in 1991 and served as the primary recruitment phase for the study. A total of 10,263 individuals of age 65 or higher were sampled at random from more than 36 communities, both rural and urban, across Canada, with specific subsets drawn from both the community and institutions for the elderly. All ten provinces of Canada were represented in the sampling procedure. These individuals were assessed for various demential conditions by the staff of the 18 participating field centers and followed-up according to certain guidelines. The second and third stages of the study consisted primarily of the reassessment, at five and ten-year marks, of individuals recruited in CSHA-1. Further study design details are provided in McDowell *et al.* (1994) and McDowell *et al.* (2005b). Our focus resides in using the CSHA-1 data to infer about the incidence of dementia in the Canadian population. Using the methodology developed in this paper, we have investigated general trends in the intensity and incidence rate of dementia in the Canadian population, as well as trends amongst various subgroups. To obtain incidence

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art54

intensities and rates, we resorted to smoothing via penalized splines and linear intensity extrapolation on the unidentified region.

We present only results pertaining to gender and age-specific disease. Due to the design-induced oversampling of the elderly, we performed stratification corrections with respect to age-at-recruitment, which highlighted some groups as particularly underpresented. The obtained correction factors are provided in Table 1.

Figure 1 presents the population-averaged cumulative intensity function initiating in July 1976, that is, the estimated number of disease onsets having occurred since July 1976. Figure 2 provides the intensity functions for the general population as well as for each gender. From this last plot, we observe that, apart from variations between 1980 and 1986, the intensity functions exhibit relative constancy, with approximately 58,000 new disease cases (about 40,000 women and 18,000 men) per year in the Canadian population. Figure 3 is a plot of the age-specific incidence rate of dementia pertaining to age groups 65-74, 75-84 and 85+. As is apparent in this figure, the identifiability period in older age groups is significantly shorter than in younger age groups. As expected, a clear monotone ordering in age-specific incidence rates emerges, with higher age group experiencing greater rates of dementia. An interesting observation regards the temporal decline in the incidence rate amongst individuals with age between 65 and 74 occurring between 1981 and 1985, whereby the incidence rate droped from an original level hovering 15 cases per 1,000 person-years to about 5 cases per 1,000 person-years. Apart from a mild increase around 1988, the incidence rate amongst individuals aged 75 to 85 was approximately constant at 35 cases per 1,000 person-years. Finally, amongst the eldest elderly (individuals of more than 85 years of age), incidence rates increased from around 90 cases per 1,000 person-years in 1986 to more than

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

115 cases per 1,000 person-years in 1991. These temporal changes may very well be a reflection of temporal delaying in the age-at-onset of dementia in the Canadian population, with decreased incidence rates amongst age group 65-74, increased rates amongst age group 85+ and relative stability amongst age group 75-84 (possible due to opposing effects of disease delay on the intermediate age category). It is reassuring to find that these rates closely match, at least in approximate average magnitude, the constant age-specific rates estimated from the prospective cohort followed between CSHA-1 and CSHA-2 (see CSHA (2000)).

7 Concluding remarks

The methodology presented in this paper is more flexible and robust relative to existing methodologies. The assumptions imposed in developing this methodology are rather mild. Despite this tremendous generality, under certain situations, some of these assumptions may well be violated. Chief amongst these is the Poisson structural assumption made on the onset point process. Despite the enormous flexibility provided by allowing nonparametric modelling of the intensity function of the onset point process, the assumption of independent increments should be scrutinized in given settings. In our application, this assumption posed negligible concern given the nature of demential diseases. However, in certain infectious diseases, particularly in their epidemic phases, independence between distinct onsets may be violated. Extension of the current methodology allowing for potential dependence between sampling units (i.e., irregularity of the underlying point process) will be considered in future work.

An additional point of future consideration regards the extension of the one-sample setting presented in this paper to the case of continuous covariates. Although adjustment

COBRA
Collection of Biostatistics
Research Archive

http://biostats.bepress.com/cobra/art54

for covariates can be performed rather effortlessly via semiparametric models imposed on the survival and truncation distributions directly (implying a semiparametric model on the intensity function and incidence rate), in view of parameter interpretability, it is desirable to propose explicit semiparametric models of the incidence rate. For example, under the assumption of proportionality between incidence rates for various covariate values, the model $r(t|z) = r_0(t)\exp(\beta z)$ may be considered, with the desirable property that $\exp(\beta)$ is a relative risk of disease. Such semiparametric extensions of this paper are the focus of ongoing research.

Finally, it is certainly of interest to also make use of data emanating from CSHA-2 and CSHA-3 to provide estimates of the incidence rate up to 2001 (rather than 1991). The truncation occurring at these points is not usual in that the follow-up pool (i.e., both truncated and untruncated individuals) is known in advance. Onset data are however available for untruncated individuals alone. Extensions accounting for these additional complexities are being currently being studied.

Age-at-onset			**Gender**		**Overall**
65-74	75-84	85+	Men	Women	—
1.417	1.059	0.930	1.150	1.051	1.078

Table 1: Correction factors for stratified sampling.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

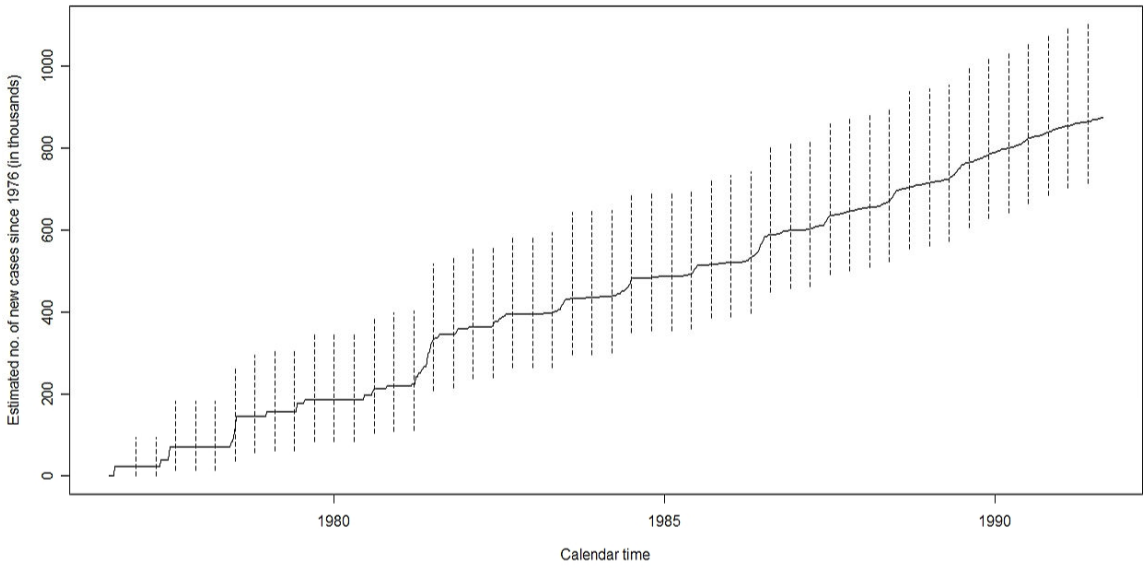

Figure 1: Estimated no. of onsets of dementia in the Canadian elderly population since 1976.

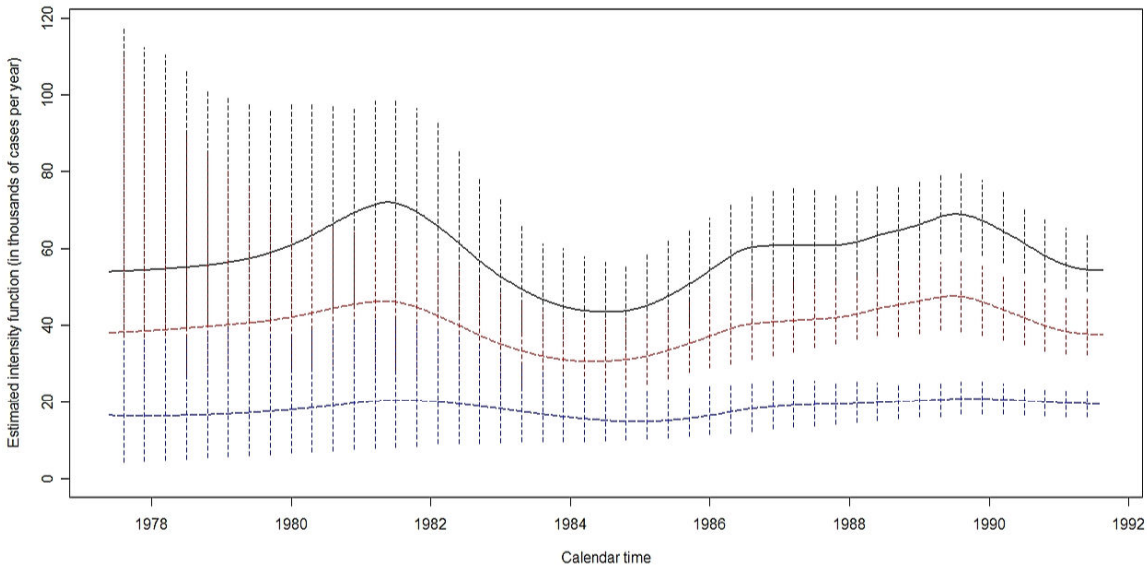

Figure 2: Estimated annual no. of onsets of dementia in the Canadian elderly population by gender (black: overall, red: women, blue: men).

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

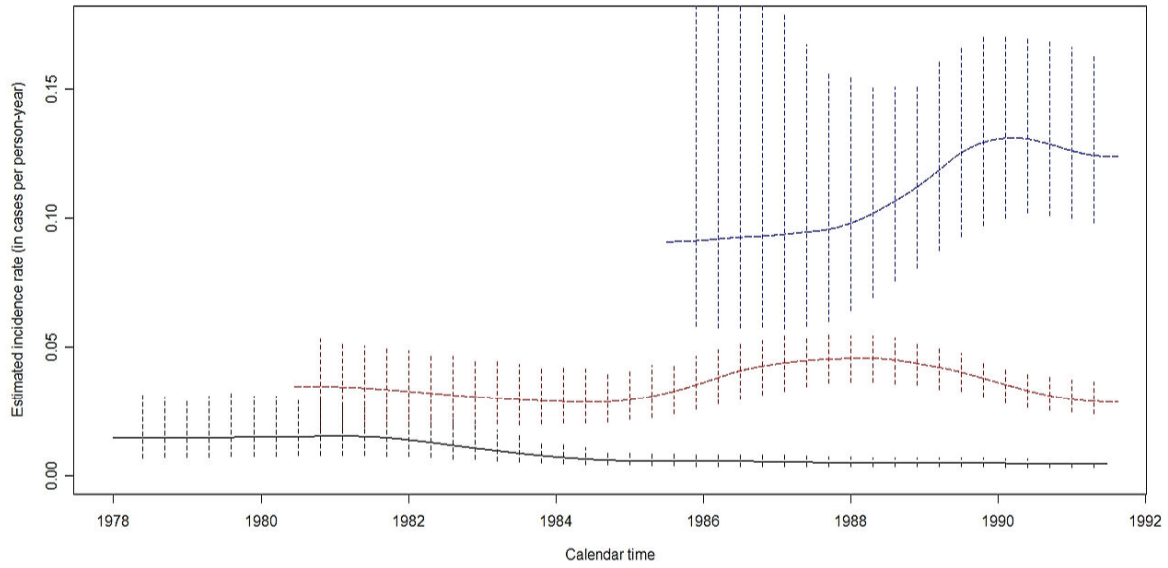

Figure 3: Estimated annual rates of dementia in the Canadian elderly population by age group (black: 65-74, red: 75-84, blue: 85+).

Acknowledgements. This work was partially supported by the National Science and Engineering Research Council of Canada and the Johns Hopkins Sommer Scholars Program. The authors thank Professors Daniel O. Scharfstein and Thomas A. Louis for helpful comments.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

8 Appendix: proofs and technical details.

Note. Left-truncation and the onset process intensity function.

Using the notation introduced earlier, the process $\{N^*(t) \equiv N(\Lambda^{-1}(t)),\ t \geq 0\}$ is a stationary Poisson process with unit rate. We may then use Theorem 2.3.1 of Ross (1983) and the Probability Integral Transform to verify that, given the occurrence of n events in $(0, \tau)$, the occurrence times $T_1, T_2, ..., T_n$ of the process $\{N(t),\ t \geq 0\}$ on $(0, \tau)$ are independent with distribution function $\Lambda(t)/\Lambda(\tau)$. The reverse-occurrence times $R_1 = \tau - T_1, R_2 = \tau - T_2, ..., R_n = \tau - T_n$ therefore have density function $\lambda(\tau - t)/\int_0^\tau \lambda(u)du$.

Proof of Theorem 1. Uniform strong consistency of $\hat{\Lambda}_c(t)$.

Denote the product space $[0, 1] \times D(0, 1) \times D(0, 1)$ by $\mathcal{D}$ and define the operator Γ acting on $\mathcal{D} \times D[0, 1] \times D[0, 1]$ as

$$\begin{aligned}\Gamma(\alpha, f_1, g_1, f_2, g_2)(t) &= \frac{\int_0^\infty S(u)dG(u)\left\{\alpha\left[1 - g_2(\tau - t)\right] - g_1(\tau - t)P_\tau\right\}}{\int_0^\infty S(u)dG(u)\int_0^\infty f_2(u)dg_2(u)} \\ &\quad - \frac{P_\tau\left[1 - G(\tau - t)\right]\left\{\int_0^\infty f_1(u)dG(u) - \int_0^\infty g_1(u)df_2(u)\right\}}{\int_0^\infty S(u)dG(u)\int_0^\infty f_2(u)dg_2(u)}\ .\end{aligned}$$

Define $\mathbb{H}_{n_1,n_2}(u)$ to be the empirical process $\sqrt{n_1}(\hat{H}_{n_2}(u) - H(u))$ for some estimator $\hat{H}_n(u)$ of $H(u)$ based on n observations, and write $\mathbb{H}_n(u)$ for $\mathbb{H}_{n,n}(u)$. This notation is needed to incorporate the fact that the index n_d of the involved estimators is random. This problem will be minor however since $n_d/n_s \xrightarrow{P} P_\tau$. By arithmetic expansion, we may verify that $\mathbb{L}_{n_s}(t) = \Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t)$ for each $t < \tau$. Now, define the operator Γ_0 acting on $\mathcal{D}$ as $\Gamma_0(\alpha, f, g)(t) \equiv \Gamma(\alpha, f, g, S, G)(t)$. Then, we have that

$$\mathbb{L}_{n_s}(t) = \Gamma_0(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d})(t) + o_p(1)\ . \tag{17}$$

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

To verify this statement, it suffices to show that

$$\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t) - \Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, S, G)(t) \xrightarrow{P} 0$$

holds uniformly in t. We may write

$$\begin{aligned}
&|\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t) - \Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, S, G)(t)| \\
&\leq \left|\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t) - \frac{\int_0^\infty \hat{S}_{n_d}(u)d\hat{G}_{n_d}(u)}{\int_0^\infty S(u)dG(u)}\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t)\right| \\
&\quad + \left|\frac{\int_0^\infty \hat{S}_{n_d}(u)d\hat{G}_{n_d}(u)}{\int_0^\infty S(u)dG(u)}\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t) - \Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, S, G)(t)\right| \\
&\leq \left|\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t)\right| \left|1 - \frac{\int_0^\infty \hat{S}_{n_d}(u)d\hat{G}_{n_d}(u)}{\int_0^\infty S(u)dG(u)}\right| \\
&\quad + \left|\frac{\mathbb{P}_{n_s}}{\int_0^\infty S(u)dG(u)}\right| \left|\hat{G}_{n_d}(\tau - t) - G(\tau - t)\right| \\
&\quad + P_\tau(1 - G(\tau - t)) \left(\int_0^\infty S(u)dG(u)\right)^{-2} \left|\int_0^\infty \mathbb{G}_{n_s,n_d}(u)d(\hat{S}_{n_d} - S)(u)\right| \\
&\leq \left|\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t)\right| \left|1 - \frac{\int_0^\infty \hat{S}_{n_d}(u)d\hat{G}_{n_d}(u)}{\int_0^\infty S(u)dG(u)}\right| \\
&\quad + \frac{n_{pop}}{\sqrt{n_s}} \left(\int_0^\infty S(u)dG(u)\right)^{-2} \left\{|\mathbb{P}_{n_s}\mathbb{G}_{n_s,n_d}(\tau - t)| \int_0^\infty S(u)dG(u) + \left|\int_0^\infty \mathbb{G}_{n_s,n_d}(u)d\mathbb{S}_{n_s,n_d}(u)\right|\right\},
\end{aligned}$$

and thus, we find that

$$\begin{aligned}
&\sup_t \left|\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G})(t) - \Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_S,n_d}, S, G)(t)\right| \\
&\leq \left|1 - \frac{\int_0^\infty \hat{S}_{n_d}(u)d\hat{G}_{n_d}(u)}{\int_0^\infty S(u)dG(u)}\right| \sup_t \left|\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t)\right| \\
&\quad + \frac{1}{\sqrt{n_s}} \left(\int_0^\infty S(u)dG(u)\right)^{-2} \sup_t \left\{|\mathbb{P}_{n_s}\mathbb{G}_{n_s,n_d}(\tau - t)| \int_0^\infty S(u)dG(u) + \left|\int_0^\infty \mathbb{G}_{n_s,n_d}(u)d\mathbb{S}_{n_s,n_d}(u)\right|\right\}
\end{aligned}$$

Upon inspection, we note that each of

$$\Gamma(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d}, \hat{S}_{n_d}, \hat{G}_{n_d})(t) \quad \text{and}$$

COBRA
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

$$\left\{|\mathbb{P}_{n_s}\mathbb{G}_{n_s,n_d}(\tau-t)|\int_0^\infty S(u)dG(u)+\left|\int_0^\infty \mathbb{G}_{n_s,n_d}(u)d\mathbb{S}_{n_s,n_d}(u)\right|\right\}$$

converge weakly (as processes) to some non-degenerate laws. Thus, since by the uniform consistency of $\hat{S}_{n_d}$ and $\hat{G}_{n_d}$ (see Tsai *et al.* (1987) and Wang (1991)) the factors

$$1-\frac{\int_0^\infty \hat{S}_{n_d}(u)d\hat{G}_{n_d}(u)}{\int_0^\infty S(u)dG(u)} \quad \text{and} \quad \frac{1}{\sqrt{n_s}}$$

converge to zero, we have verified that

$$\sup_t \left|\Gamma(\mathbb{P}_{n_s},\mathbb{S}_{n_s,n_d},\mathbb{G}_{n_s,n_d},\hat{S}_{n_d},\hat{G}_{n_d})(t)-\Gamma(\mathbb{P}_{n_s},\mathbb{S}_{n_s,n_d},\mathbb{G}_{n_s,n_d},S,G)(t)\right| \xrightarrow{P} 0 \ .$$

It is not difficult to verify, additionally, that the operator Γ_0 is bounded and linear, and this fact is crucial in our approach to determining the asymptotic law of $\mathbb{L}_{n_s}(t)$. Indeed, linearity and boundedness jointly suffice to ensure the continuity of the operator Γ_0.

To verify the uniform consistency of $\hat{\Lambda}(t)$, we use (17) and write

$$\begin{aligned}
\hat{\Lambda}(t)-\Lambda(t) &= \frac{1}{\sqrt{n_s}}\mathbb{L}_{n_s}(t) = \frac{1}{\sqrt{n_s}}\Gamma_0\left(\mathbb{P}_{n_s},\mathbb{S}_{n_s,n_d},\mathbb{G}_{n_s,n_d}\right)(t)+o_p\left(\frac{1}{\sqrt{n_s}}\right)\\
&= \Gamma_0\left(\frac{1}{\sqrt{n_s}}\mathbb{P}_{n_s},\frac{1}{\sqrt{n_s}}\mathbb{S}_{n_s,n_d},\frac{1}{\sqrt{n_s}}\mathbb{G}_{n_s,n_d}\right)(t)+o_p\left(\frac{1}{\sqrt{n_s}}\right)\\
&= \Gamma_0\left(\hat{P}_{n_s}-P,\hat{S}_{n_d}-S,\hat{G}_{n_d}-G\right)(t)+o_p\left(\frac{1}{\sqrt{n_s}}\right)\\
&= \Gamma_0(\hat{P}_{n_s},\hat{S}_{n_d},\hat{G}_{n_d})(t)-\Gamma_0\left(P,S,G\right)(t)+o_p\left(\frac{1}{\sqrt{n_s}}\right) \ .
\end{aligned}$$

Since Γ_0 is a continuous operator and that $(\hat{P},\hat{S},\hat{G})$ is uniformly consistent for (P,S,G), we have that

$$\Gamma_0(\hat{P}_{n_s},\hat{S}_{n_d},\hat{G}_{n_d})(t)-\Gamma_0\left(P,S,G\right)(t)\xrightarrow{P}0$$

uniformly in t, and so, we conclude that $\hat{\Lambda}$ is uniformly consistent for $\Lambda(t)$.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics

Proof of Theorem 2. Weak convergence of $\sqrt{n_s}(\hat{\Lambda}_c(t)-\Lambda_c(t))$.

The centralized variable $\mathbb{P}_{n_s}$ is asymptotically normal by the usual CLT for independent ran-

http://biostats.bepress.com/cobra/art54

dom variables while the joint process $(\mathbb{S}_{n_d}, \mathbb{G}_{n_d})$ is asymptotically Gaussian, as shown by Wang (1991). The (asymptotic) independence of $\mathbb{P}_{n_s}$ and $(\mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d})$ (see Addona *et al.* (2009)) suffices then to establish the asymptotic convergence of $(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d})$ to a Gaussian process, which, from the consistency of the involved estimators, has mean zero. By the representation above, the asymptotic distribution of the process $\mathbb{L}_{n_s}(t)$ is simply that of the process $\Gamma_0(\mathbb{P}_{n_s}, \mathbb{S}_{n_s,n_d}, \mathbb{G}_{n_s,n_d})$, and since Γ_0 is continuous, the Extended Continuous Mapping Theorem, as stated in Kosorok (2008), applies. By the linearity of Γ_0, the asymptotic distribution of $\mathbb{L}_{n_s}(t)$ is Gaussian with mean zero and covariance function $\Sigma(s,t)$.

Proof of Theorem 3. Uniform strong consistency of $\hat{R}_c(t)$.

Similar to the proof of Theorem 1. Defining the operator

$$\Upsilon(f)(t) \;=\; \int_0^t \frac{df(u)}{Q(u) - \Lambda_u^\star(u)} \;,$$

the result follows from Theorem 1 and the representation

$$\sqrt{n_s}(\hat{R}_c(t) - R_c(t)) \;=\; \Upsilon\left(\sqrt{n_s}(\hat{\Lambda}_c - \Lambda_c)\right)(t) \;,$$

with Υ easily verified to be both linear and bounded.

Proof of Theorem 4. Weak convergence of $\sqrt{n_s}(\hat{R}_c(t) - R_c(t))$.

Similar to the proof of Theorem 2, using Theorem 2 and the representation discussed above.

COBRA
A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press

References

Addona, V., Asgharian, M., & Wolfson, D.B. 2009. On the incidence-prevalence relation and length-biased sampling. *Canadian Journal of Statistics*, **37**(2), ?–?

Albert, A., Gertman, PM, & Louis, TA. 1978a. Screening for the early detection of cancer. I. The temporal natural history of a progressive disease state. *Mathematical Biosciences*, **40**, 1–59.

Albert, A., Gertman, PM, Louis, TA, & Liu, SI. 1978b. Screening for the early detection of cancer. II. The impact of screening on the natural history of the disease. *Mathematical Biosciences*, **40**, 61–109.

Alho, JM. 1992. On prevalence, incidence, and duration in general stable populations. *Biometrics*, **48**(2), 587–592.

Alioum, A., Commenges, D., Thiebaut, R., & Dabis, F. 2005. A multistate approach for estimating the incidence of human immunodeficiency virus by using data from a prevalent cohort study. *Journal of the Royal Statistical Society Series C*, **54**(4), 739–752.

Asgharian, M., M'Lan, C.E., & Wolfson, D.B. 2002. Length-biased sampling with right-censoring: an unconditional approach. *Journal of the American Statistical Association*, **97**(457), 201–210.

Brookmeyer, R., & Damiano, A. 1989. Statistical methods for short-term projections of AIDS incidence. *Statistics in Medicine*, **8**(1), 23–34.

Brookmeyer, R., & Quinn, T.C. 1995. Estimation of Current Human Immunodeficiency Virus Incidence Rates from a Cross-Sectional Survey Using Early Diagnostic Tests. *American Journal of Epidemiology*, **141**(2), 166–172.

Cheng, MY, Hall, P., & Yang, Y.J. 2007. Nonparametric inference under dependent truncation. *Acta Scientiarum Mathematicarum*, **73**(1-2), 397–422.

Cox, D.R., & Oakes, D. 1984. *Analysis of Survival Data*. Chapman & Hall/CRC.

CSHA. 2000. The Incidence of Dementia in Canada. *Neurology*, **55**(1), 66–73.

Daley, D.J., & Vere-Jones, D. 2003. *An Introduction to the Theory of Point Processes*. Springer.

Freeman, J., & Hutchison, G.B. 1980. Prevalence, Incidence and Duration. *American Journal of Epidemiology*, **112**(5), 707–723.

Keiding, N. 1991. Age-Specific Incidence and Prevalence: A Statistical Perspective. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **154**(3), 371–412.

Keiding, N., Holst, C., & Green, A. 1989. Rretrospective estimation of diabetes incidence from information in a prevalent population and historical mortality. *American Journal of Epidemiology*, **130**(3), 588–600.

http://biostats.bepress.com/cobra/art54

Kosorok, M.R. 2008. *Introduction to Empirical Processes and Semiparametric Inference*. Springer Verlag.

Louis, TA, Albert, A., & Heghinian, S. 1978. Screening for the early detection of cancer. III. Estimation of disease natural history. *Mathematical Biosciences*, **40**, 111–144.

McDowell, I., Hill, G., Lindsay, J., Helliwell, B., Costa, L., Beattie, B., Tuokko, H., Hertzman, C., Gutman, G., & Parhad, I. 1994. Canadian Study of Health and Aging: Study Methods and Prevalence of Dementia. *Canadian Medical Association Journal*, **150**, 899–912.

McDowell, I., Hill, G., & Lindsay, J. 2005a. An Overview of the Canadian Study of Health and Aging. *International Psychogeriatrics*, **13**(S1), 7–18.

McDowell, I., Aylesworth, R., Stewart, M., Hill, G., & Lindsay, J. 2005b. Study Sampling in the Canadian Study of Health and Aging. *International Psychogeriatrics*, **13**(S1), 19–28.

Miettinen, O. 1976. Estimability and Estimation in Case-Referent Studies. *American Journal of Epidemiology*, **103**(2), 226–235.

Ramlau-Hansen, H. 1983. Smoothing Counting Process Intensities by Means of Kernel Functions. *The Annals of Statistics*, **11**(2), 453–466.

Ross, S.M. 1983. *Stochastic processes*. Wiley New York.

Rothman, KJ, Greenland, S., & Lash, TL. 2008. *Modern Epidemiology*. Philadelphia, USA: Lippincott Williams & Wilkins.

Schabenberger, O., & Gotway, C.A. 2005. *Statistical Methods For Spatial Data Analysis*. CRC Press.

Szklo, M., & Nieto, F. 2000. *Epidemiology: beyond the basics.* MD: Aspen Publishers, Inc.

Tsai, W., Jewell, N.P., & Wang, M.C. 1987. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, **74**(4), 883–886.

Wang, M.C. 1991. Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, **86**(413), 130–143.

Wang, MC. 1998. Length bias. *Encyclopedia of Biostatistics*, **3**, 2223–2226.

Wang, M.C., Brookmeyer, R., & Jewell, N.P. 1993. Statistical models for prevalent cohort data. *Biometrics*, **49**(1), 1–11.

Wolfson, C., Wolfson, D.B., Asgharian, M., M'Lan, C.E., Ostbye, T., Rockwood, K., Hogan, DB, *et al.* 2001. A Reevaluation of the Duration of Survival after the Onset of Dementia. *New England Journal of Medicine*, **344**(15), 1111.

A BEPRESS REPOSITORY
Collection of Biostatistics
Research Archive

Hosted by The Berkeley Electronic Press