Nonparametric incidence estimation from prevalent cohort survival data

BY MARCO CARONE

*Division of Biostatistics, School of Public Health, University of California Berkeley,
109 Haviland Hall, Berkeley, California 94720, U.S.A.
mcarone@berkeley.edu*

MASOUD ASGHARIAN

*Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West,
Montreal, Quebec H3A 2K6, Canada
masoud@math.mcgill.ca*

AND MEI-CHENG WANG

*Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University,
615 North Wolfe Street, Baltimore, Maryland 21205, U.S.A.
mcwang@jhsph.edu*

SUMMARY

Incidence is an important epidemiological concept most suitably studied using an incident cohort study. However, data are often collected from the more feasible prevalent cohort study, whereby diseased individuals are recruited through a cross-sectional survey and followed in time. In the absence of temporal trends in survival, we derive an efficient nonparametric estimator of the cumulative incidence based on such data and study its asymptotic properties. Arbitrary calendar time variations in disease incidence are allowed. Age-specific incidence and adjustments for both stratified sampling and temporal variations in survival are also discussed. Simulation results are presented and data from the Canadian Study of Health and Aging are analysed to infer the incidence of dementia in the Canadian elderly population.

Some key words: Age-specific incidence; Cross-sectional sampling; Left-truncation; Point process; Stratification.

1. INTRODUCTION

To study the characteristics of some disease or condition of interest in a population, 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. The incident cohort study is considered to be the gold standard in inferring disease incidence in a target population. But it can be unfeasible to conduct, since extensive enrolment and long follow-up periods are often required to accrue a sufficient number of disease cases, particularly in relatively rare diseases (Rothman et al., 2008, p. 108).

As an alternative, investigators sometimes conduct prevalent cohort studies by recruiting diseased individuals at some fixed time, determining their onset time retrospectively and

following them until death or loss to follow-up. This design is often easier to conduct, although the data it generates incorporate systematic biases that, if ignored, can lead to grossly incorrect conclusions. [Wolfson et al. \(2001\)](#) highlight an example of this. In particular, individuals with longer disease durations are overrepresented by the sampling scheme. This bias in survival has been widely studied ([Cox & Oakes, 1984](#), p. 179; [Tsai et al., 1987](#); [Wang, 1991](#); [Asgharian et al., 2002](#)).

If interest lies in disease incidence, it is not a priori clear whether data from a prevalent cohort study suffice to identify the underlying incidence process. At least two issues complicate inference regarding incidence. The first is the absence of a clearly identified pool of disease-free individuals, susceptible to onset of disease. In the incident cohort study, all individuals are initially disease-free; as time progresses, however, some become diseased. Incidence can be gauged by comparing the number of incident cases relative to the size of the susceptible pool and projecting back to the whole population. In the prevalent cohort study, since all individuals are diseased upon recruitment, such direct comparisons using available risk pools are not possible. The second is the overrepresentation of individuals with recent onsets in the sample, due to the fact that individuals with onsets in the distant past are less likely to survive until the time of recruitment. Naive analyses ignoring this bias may suggest increasing rates of disease incidence over calendar time even when such rates have been decreasing. This paper is concerned with the nonparametric estimation of possibly time-varying incidence rates from prevalent cohort survival data. Interest in this problem arose from an investigation into the Canadian Study of Health and Aging, a study of dementia in the Canadian elderly population.

Variants of this problem have been studied in the statistical and epidemiological literature. 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 to exponentially varying stable populations. [Keiding \(1991\)](#) provided a comprehensive look at incidence estimation from current-duration data under a Markovian structure, although the absence of follow-up required the imposition of assumptions possibly difficult to verify in practice. [Keiding et al. \(1989\)](#) also considered current-duration data and suggested the use of inverse-weighting using background knowledge of historical mortality. [Brookmeyer & Quinn \(1995\)](#) focused on HIV infection incidence rate estimation from cross-sectional data, and [Alioum et al. \(2005\)](#) studied estimation of HIV infection incidence from prevalent cohort data using multistate Markov models. More recently, [Addona et al. \(2009\)](#) provided inferential results on incidence estimation under equilibrium conditions, whereby both population size and the number of disease onsets per unit of time are assumed to be constant through time. We propose a flexible and efficient nonparametric estimation framework allowing for arbitrary calendar time variations in disease incidence and temporal changes in population size. We discuss how our methodology may be used to estimate age-specific rates, and how to adjust for stratification in the sampling scheme as well as potential temporal trends in disease duration. The methodology proposed is an example of a deconvolution problem, somewhat in the spirit of the back-calculation method elaborated by [Brookmeyer & Damiano \(1989\)](#) in the setting of HIV infection and the approach of [Esbjerg et al. \(1999\)](#) in the study of multiple sclerosis.

2. INCIDENCE RATE

2.1. *Definition*

The incidence of a disease refers to its occurrence in a susceptible population. Let disease onset be a clearly defined event, and for simplicity, assume that the disease of interest is irreversible.

Suppose that in the target population onsets of disease arise from a nonstationary Poisson process $N = \{N(t)\}_{-\infty < t < \infty}$, the so-called onset point process, with rate function $\lambda(t)$. The indexing variable t represents calendar time.

The incidence rate, an absolute instantaneous measure of disease occurrence, is the expected number of new disease cases per unit of time in a prespecified population. Probabilistically, this measure is simply the rate function $\lambda(t)$ of the onset point process. Its use is particularly important in public health, where planning and allocation of health resources require knowledge of the magnitude of disease burden over time.

2.2. Probabilistic framework

Suppose that sampling occurs at a fixed time τ . Define disease duration X^0 as the time elapsed between disease onset and death, and truncation time T^0 as the time elapsed between disease onset and recruitment. All discussions are restricted to the subpopulation of individuals experiencing the disease of interest before τ , that is, individuals with $T^0 > 0$. Denote by S the survival function of X^0 , by $b < \infty$ the upper limit of the support of X^0 and by G the distribution function of T^0 given $0 < T^0 \leq b$. A key assumption made here is that disease duration is independent of calendar time of onset; more concisely, X^0 and T^0 are assumed to be independent given $0 < T^0 \leq b$. This assumption is discussed further in §3.4. Individuals have prevalent disease at τ if and only if their disease duration X^0 exceeds their truncation time T^0 . Denote by X and T the counterparts of X^0 and T^0 , respectively, in the subpopulation of individuals with prevalent disease at τ . The distribution of (X, T) is that of (X^0, T^0) conditional upon $X^0 > T^0 > 0$. Suppose that for prevalent individuals the residual disease duration $X - T$ is potentially right-censored by a residual censoring variable $C > 0$; denote the observed disease duration and event indicator by $Y = \min(X, T + C)$ and $\Delta = \mathbb{I}_{Y=X}$, respectively. Censoring applies to the residual disease duration because it is study dependent and acts exclusively upon a subject's time under study. As in Wang (1991) and Asgharian et al. (2002), the residual censoring mechanism is assumed to be uninformative: specifically, $(T, X - T)$ and C are assumed to be independent. In many applications, censoring occurs for administrative or other reasons unrelated to participants' history; this assumption is adequate in such situations.

Define the cumulative rate function $\Lambda(t) = \int_{\tau-b}^t \lambda(u) du$ of the onset point process, hereafter referred to as the cumulative incidence, with value 0 for $t \leq \tau - b$. The cumulative incidence is interpreted as follows: for $\tau - b < t_1 < t_2$, the difference $\Lambda(t_2) - \Lambda(t_1)$ is the mean number of disease onsets occurring in the target population between calendar times t_1 and t_2 . Define the τ -observable onset point process $M = \{M(t)\}_{-\infty < t \leq \tau}$ such that $M(t)$ is the number of individuals with disease onset before calendar time t , both alive and diseased at τ . A jump in process N at time u can result in a jump in process M at time u or no jump at all; the former event occurs with probability $\text{pr}(X^0 \geq \tau - u \mid T^0 = \tau - u)$. In view of the independent truncation assumption, M is a thinning of N with thinning function $p(u) = S(\tau - u)$. It follows that M is a nonstationary Poisson process with the cumulative rate function

$$\Phi(t) = \int_{-\infty}^t S(\tau - u) \lambda(u) du = \int_{-\infty}^t S(\tau - u) d\Lambda(u). \quad (1)$$

As expected, since individuals with onsets before $\tau - b$ cannot be observed at τ , $\Phi(t) = 0$ for $t \leq \tau - b$.

2.3. Identifiability

The problem consists in deconvolving (1) in order to draw inference about $\Lambda(\cdot)$ using right-censored disease duration data on a sample from M . The duration data from prevalent individuals sampled at recruitment time identify $\lambda(t)$ up to a constant of proportionality on $(\tau - b, \tau)$. This follows from the fact that, for $t \leq \tau$,

$$\Lambda(t) = \kappa \{1 - G(\tau - t)\} \quad (2)$$

for some constant $\kappa > 0$, proved in Appendix A, and that the truncation distribution G is identified (Wang, 1991). 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 as well. This fact follows from the combination of (1) and (2) into

$$\Phi(t) = \kappa \int_{\tau-t}^{\infty} S(u) dG(u),$$

and the observation that $\Phi(\tau)$ is the mean population prevalence at recruitment. Thus, the cumulative incidence function may be written as

$$\Lambda(t) = \frac{\Phi(\tau)}{\int_0^{\infty} S(u) dG(u)} \{1 - G(\tau - t)\} \quad (3)$$

$$= \Phi(\tau) \int_{\tau-t}^{\infty} \frac{dG'(u)}{S(u)}, \quad (4)$$

where G' is the distribution function of the observable truncation time T , that is, the distribution function of T^0 given $X^0 > T^0 > 0$. It is possible to derive (4) directly from (1) rather than as a consequence of (3). Manipulation of (1) yields that

$$\Lambda(t) = \int_{-\infty}^t \frac{d\Phi(u)}{S(\tau - u)}$$

for each $t \leq \tau$. Using an argument similar to that in Appendix A, it can also be shown that $\Phi(t) = \Phi(\tau)\{1 - G'(\tau - t)\}$. Together, these two equations lead to (4).

In order to obtain more tractable relationships, epidemiologists have often considered the so-called stable disease conditions, which assume a stationary onset point process and stable population size. Together, these requirements imply both that disease prevalence is constant over time and that G is the distribution function of a uniform variate over $[0, b)$, and (3) then reduces to prevalence = incidence $\times$ mean duration, a well-known result in classical epidemiology (Miettinen, 1976; Keiding, 1991). The framework provided in this paper may thus be seen as a generalization of this result, allowing for arbitrarily varying incidence and population size 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 at recruitment time and each recruit is assessed for prevalent disease. Thus, in addition to providing a sample of prevalent individuals, each contributing onset and survival data, an estimate of population disease prevalence is also obtained. Identifiability of the incidence rate $\lambda(\cdot)$ on $(\tau - b, \tau)$ follows. The weakening of this assumption is considered in § 3.

2.4. Estimation

The data consist of n_d independent triplets (T_i, Y_i, Δ_i) , where n_d is the number of prevalent cases among n_s sampled individuals. Suppose that n is the size of the target population, that is, the number of individuals alive and satisfying the study's inclusion criteria at any time between $\tau - b$ and τ , and let n_{pop} be the size of the population at recruitment. From the discussion above, for $\tau - b < t \leq \tau$, a natural substitution estimator for $\Lambda(t)$ emerges as

$$\hat{\Lambda}(t) = \hat{\Phi}(\tau) \int_{\tau-t}^{\infty} \frac{d\hat{G}'(u)}{\hat{S}(u)}.$$

In the above, $\hat{\Phi}(\tau)$ is the estimated population prevalence $n_{\text{pop}}\hat{\phi}(\tau)$, where $\hat{\phi}(\tau) = n_d/n_s$ is an estimator of the mean population prevalence proportion at τ , $\phi(\tau)$, and $\hat{S}$ is the truncation product-limit estimator (Kaplan & Meier, 1958; Tsai et al., 1987):

$$\hat{S}(u) = \prod_{i=1}^n \left\{ 1 - \frac{\Delta L(Y_i)}{R(Y_i)} \right\}^{L_i(u)\Delta_i},$$

where $L_i(u) = \mathbb{I}_{(-\infty, Y_i]}(u)$, $L(u) = \sum_{i=1}^n \Delta_i L_i(u)$ and $R(u) = \sum_{i=1}^n \mathbb{I}_{[T_i, Y_i]}(u)$. Finally, $\hat{G}'$ is the empirical distribution function based on observed truncation times. It follows that

$$\hat{\Lambda}(t) = \frac{n_{\text{pop}}}{n_s} \sum_{i=1}^{n_d} \frac{\mathbb{I}_{[0, T_i]}(\tau - t)}{\hat{S}(T_i)}.$$

In practice, although all data should be used to estimate each of $\hat{\Phi}(\tau)$, $\hat{S}$ and $\hat{G}'$, it is generally advantageous to consider estimating a trimmed version of $\Lambda(t)$, say $\Lambda(t; c) = \Lambda(t) - \Lambda(c) = \int_c^t \lambda(u) du$ for some $c > \tau - b$ chosen to discard the earliest onset points. This consideration alleviates the problem of excess variability due to observed onsets particularly distant from recruitment time and therefore overly influential on the final estimate. Increments of $\Lambda(t)$ and $\Lambda(t; c)$ have identical interpretations, and the natural substitution estimator for $\Lambda(t; c)$ is

$$\hat{\Lambda}(t; c) = \frac{n_{\text{pop}}}{n_s} \sum_{i=1}^{n_d} \frac{\mathbb{I}_{[0, T_i]}(\tau - t) \mathbb{I}_{(T_i, \infty]}(\tau - c)}{\hat{S}(T_i)}.$$

2.5. Asymptotic properties

Because the sample size n_s cannot exceed the size n_{pop} of the cross-sectional population, which itself cannot exceed the size n of the target population, both n_s and n are sent to infinity when studying the asymptotic properties of the proposed estimator. Thus, both the estimator and estimand considered depend on and grow with n . Certain properties of the estimator $\hat{\Lambda}(t; c)$ are presented in Theorem 1. A sketch of the proof is outlined in Appendices B and C, and further technical details are provided in the Supplementary Material. Below, D_{0c} denotes the space of cadlag functions on $[c, \tau]$ vanishing at c .

THEOREM 1. *If S and G are continuous, then $\sup_{t \in [c, \tau]} |\hat{\Lambda}(t; c) - \Lambda(t; c)|/n_{\text{pop}}$ converges to zero in probability as n and n_s tend to ∞ for any $\tau - b \leq c < \tau$. Furthermore, if $\int_0^\infty d\{1 - S(u)\}/G(u) < \infty$ and $n_s/n \rightarrow 0$ as n and n_s tend to ∞ , then, for any $\tau - b < c < \tau$,*

the process

$$\hat{A}(t; c) = n_s^{1/2} \left\{ \frac{\hat{\Lambda}(t; c) - \Lambda(t; c)}{n_{\text{pop}}} \right\}$$

converges weakly, as an element of D_{0c} in the Skorohod topology, to a mean zero Gaussian process with covariance function Σ_c given in Appendix C, as n and n_s tend to ∞ .

The integral condition requires the disease duration distribution to have a sufficiently light left tail. Provided the incidence rate is continuous and strictly positive in a neighbourhood of τ , this condition will be satisfied if, for example, there exists some $\alpha > 1$ such that $\lim_{u \rightarrow 0+} \{1 - S(u)\}/u^\alpha < \infty$. The condition $n_s/n \rightarrow 0$ requires that the sample size grow at a slower rate than the size of the target population. Habitual resource limitations render this plausible in practice. In the case $n_s/n \rightarrow \nu \in (0, 1]$, the proposed estimator also behaves as a Gaussian process asymptotically, although its asymptotic covariance function differs from Σ_c by a component that cannot be estimated from the available data. Since each of $\hat{\Phi}(\tau)$, $\hat{S}$ and $\hat{G}'$ is efficient for their respective targets, $\hat{\Lambda}(t; c)$ is itself efficient for $\Lambda(t; c)$ (van der Vaart, 2000, Theorem 25.47). While empirical estimates may be substituted in Σ_c to form confidence intervals and bands for $\hat{\Lambda}(t)$, it may be more expedient to employ resampling techniques. The bootstrap here must be performed in two stages, reflecting the data generation process. First, the number of prevalent individuals sampled n_d^{boot} is generated from a binomial distribution with size n_s and success probability n_d/n_s . Secondly, the survival data of the n_d prevalent individuals are resampled with replacement to obtain a sample of n_d^{boot} survival triplets. The bootstrap sample obtained then comprises the generated size n_d^{boot} of the disease subsample and the resampled survival data. We conjecture that quantile-based confidence intervals based on this bootstrap scheme provide intervals with adequate empirical coverage, and simulation studies presented in §4 indeed support this claim. A formal proof of validity may possibly build upon the validity of the bootstrap for each of $\hat{\Phi}(\tau)$, $\hat{S}$ and $\hat{G}'$, the asymptotic independence of these estimators, discussed in Appendix C, and the delta method for the bootstrap (van der Vaart & Wellner, 1996, Theorem 3.9.11).

It may be of interest to test whether the incidence rate has been constant during the period of identifiability. Such a test may certainly be constructed from the estimator $\hat{\Lambda}(t; c)$. However, a direct consequence of (2) is that the incidence rate is constant if and only if the truncation distribution G is uniform; a test of stationarity of the onset point process may thus be based on $\hat{G}$ alone. Addona & Wolfson (2006) devise such a test.

3. APPLICATIONS AND EXTENSIONS

3.1. Standardized incidence rate

The scale of the incidence rate depends on the size of the underlying population and thus cannot be directly used to compare disease risk in subpopulations. For example, a low-risk subpopulation may have the same incidence rate as a high-risk subpopulation if, say, the former is much larger in size than the latter. Furthermore, secular variations in the incidence rate may be due to changes in population size or inherent disease risk. As such, the standardized incidence rate $r(t)$, a relative instantaneous measure of disease occurrence, may be of interest. It is defined as

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

where $Q(t)$ and $P(t)$ are, respectively, the size of the population and the size of its prevalent subpopulation at calendar time t . Every individual in the population need not be at risk for the disease in question. Diseased individuals, although included in the 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 $Q(t)$ as has often 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, the incidence rate may be seriously underestimated without this adjustment.

The incidence rate constitutes a true measure of disease risk. Its uses include, for example, the identification of disease risk factors and the assessment of interventions aimed at disease prevention. If external information on the size of the population and the prevalent subpopulation through time is available, a straightforward estimator of the cumulative standardized incidence $R(t; c) = \int_c^t \{Q(u) - P(u)\}^{-1} d\Lambda(u)$ is obtained. The former can often be obtained from census data. The latter, however, is usually more problematic. Without historical prevalence information, some form of extrapolation based on the available data is inevitable. Under stable disease conditions, the relationship above reduces to the well-known epidemiological equation: prevalence odds = incidence rate $\times$ mean duration (Keiding, 1991; Rothman et al., 2008, p. 48).

3.2. Age-specific incidence

Estimation of population-averaged disease incidence has been discussed so far. Estimation of disease incidence for particular subpopulations, however, may easily be performed by adequate subgroup analyses. Suppose that Z^0 is a categorical covariate, and assume that within the strata of Z^0 , X^0 and T^0 are independent. Then, redefining the reference population considered, the same derivation used to obtain (4) yields that the cumulative incidence pertaining to $Z^0 = z$ is

$$\Lambda_z(t) = \Phi_z(\tau) \int_{\tau-t}^{\infty} \frac{dG'_z(u)}{S_z(u)},$$

where S_z , G'_z and $\Phi_z(\tau)$ are the stratum-specific counterparts of S , G' and $\Phi(\tau)$, respectively. Each of S_z and G'_z is estimated by considering the subset of the prevalent cohort data for which $Z = z$, where Z is the counterpart of Z^0 defined in the cross-sectional population at τ . Stratum-specific prevalence $\Phi_z(\tau)$ at time τ is estimated consistently by $n_{d,z}n_{\text{pop}}/n_s$, where $n_{d,z}$ is the number of stratum-specific prevalent individuals sampled. An estimator of $\Lambda_z(t)$ is thus

$$\hat{\Lambda}_z(t) = \frac{n_{\text{pop}}}{n_s} \sum_i \frac{\mathbb{I}(Z_i = z) \mathbb{I}_{[0, T_i]}(\tau - t)}{\hat{S}_z(T_i)},$$

where $\hat{S}_z$ is the obvious subgroup counterpart of $\hat{S}$. If Z^0 is taken to be age at disease onset, then the prevalence $\Phi_z(\tau)$ of age-specific disease should not be confused 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. Results from a simulation study illustrating the appropriateness of this methodology for estimating age-specific incidence rates are provided in the Supplementary Material.

3.3. Stratified sampling

Recruited individuals have so far been assumed to form a simple random sample from the population at time τ . In many cases, however, the sampling scheme incorporates stratification in

order to increase the recruitment rate for otherwise underrepresented subgroups of the population. For example, in the case of the Canadian Study of Health and Aging, individuals living in institutions for the elderly were oversampled by design. With knowledge of certain population characteristics, it is possible to recover correct population estimates from sample quantities.

Let U^0 be a categorical variable, defined for all individuals in the population, and denote by U the counterpart of U^0 in the cross-sectional population at τ . Suppose that U is the stratifying variable: within levels of U , observations form a simple random sample. Since each onset arises from exactly one level of U , the cumulative incidence for the subpopulation with $Z^0 = z$ may be written as

$$\Lambda_z(t) = \sum_u \Lambda_{u,z}(t) = \sum_u \Phi_{u,z}(\tau) \int_{\tau-t}^{\infty} \frac{dG'_{u,z}(w)}{S_{u,z}(w)}, \quad (5)$$

where $\Lambda_{u,z}$, $\Phi_{u,z}(\tau)$, $G'_{u,z}$ and $S_{u,z}$ are the cumulative incidence, the number of prevalent individuals at τ , the distribution function of the observable truncation time and the survival function of disease duration, respectively, within the subpopulation with $(U^0, Z^0) = (u, z)$. Substitution of empirical quantities in (5) leads to the estimator

$$\hat{\Lambda}_z(t) = \sum_u \left\{ \frac{n_{\text{pop},u}}{n_{s,u}} \sum_i \frac{\mathbb{I}(U_i = u, Z_i = z) \mathbb{I}_{[0, T_i]}(\tau - t)}{\hat{S}_{u,z}(T_i)} \right\},$$

where $n_{\text{pop},u}$ and $n_{s,u}$ are the size of the population in category $U = u$ at τ and the number of such individuals sampled, respectively, and $\hat{S}_{u,z}$ is the truncation product-limit estimator based on all disease duration data with $(U, Z) = (u, z)$. The above adjustment for stratification requires estimating the cumulative incidence within various subpopulations, necessarily increasing the variance of the population-averaged estimator. To alleviate this problem, rather than performing nonparametric estimation within levels of the stratifying variable, a practitioner may consider modelling the disease duration distribution using a semiparametric model incorporating the stratification variable as a regressor. The resulting gain in efficiency would, nonetheless, come at the expense of robustness. An example of the use of semiparametric modelling in our methodology is presented below.

3.4. Temporal trends in survival

In some applications, the independence between disease duration and onset time, or rather, the absence of any temporal trend in survival, may fail to hold. It is possible, however, to relax this assumption. According to [Cheng et al. \(2007\)](#), the onset-conditional survival distribution cannot be nonparametrically identified without specification of the dependence structure between survival and onset time. Assuming that onset and truncation times identify each other, we consider incorporating temporal trends in survival via a proportional hazards model relating survival and truncation. Specifically, denoting by $h(x | t)$ the hazard rate at time x associated with individuals with truncation time t , we consider the one-parameter regression model of survival conditional upon truncation $h(x | t) = h_0(x) \exp\{\gamma \phi(t)\}$ 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 application of interest. For example, a disease for which survival is believed to have varied only gradually over time might warrant $\phi(t) = t$ or $\phi(t) = \log t$, while a step function might be more appropriate for a disease for which a marked change in survival occurred, due, for example, to the discovery of an innovative treatment. As confirmed by simulation studies not shown here, substantial robustness to the independence assumption of § 2 may be gained from this simple

semiparametric extension of our methodology, although at the cost of a possibly greater estimator variability due to the introduced parameter. In addition to 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 are performed through usual risk-set methods accounting for truncation (Wang et al., 1993). Equation (4) still holds, with $S(u)$ replaced by $S(u | u)$, where $S(x | t) = \text{pr}(X^0 > x | T^0 = t)$ is the survival function corresponding to individuals with onset at time $\tau - t$ evaluated at x . Thus, we have the estimator

$$\hat{\Lambda}(t) = \frac{n_{\text{pop}}}{n_s} \sum_{i=1}^{n_d} \frac{\mathbb{I}_{[0, T_i]}(\tau - t)}{\hat{S}_{\hat{\gamma}}(T_i | T_i)},$$

where $\hat{S}_{\hat{\gamma}}(u | u)$ is the model-based survival function estimator. Additional regressors, such as a stratifying variable, may be included in the above proportional hazards regression model, although goodness-of-fit testing would then become crucial.

4. SIMULATION STUDY

A simulation study was conducted to evaluate the finite sample performance of the proposed methodology. Two scenarios for each of the incidence rate and the disease duration distribution were investigated. The incidence rate was considered to be either increasing over time, reflected by truncation density $g_0(t) = (0.016 - 0.00012t)\mathbb{I}_{(0,100)}(t)$, or constant over time, with truncation density $g_1(t) = \mathbb{I}_{(0,100)}(t)/100$. The disease duration was taken to be either a beta variate with density $f_0(x) \propto x^{19}(100 - x)^4\mathbb{I}_{(0,100)}(x)$, or a truncated exponential variate with density $f_1(x) \propto \exp(-x/100)\mathbb{I}_{(0,100)}(x)$. The former distribution, in conjunction with both truncation distributions considered, satisfies all assumptions in Theorem 1, while the latter does not satisfy the integral condition required for weak convergence to a Gaussian process at the regular rate. This will illustrate the necessity of this condition. The calendar time at recruitment was taken to be 100, resulting in a calendar range of identifiability of $(0, 100]$. We refer to the disease duration and truncation distribution combinations (f_0, g_0) , (f_0, g_1) , (f_1, g_0) and (f_1, g_1) as scenarios I, II, III and IV, respectively. The residual censoring variable was exponentially distributed with the mean chosen to ensure a 25% censoring proportion. The disease prevalence proportion and the size n_{pop} of the cross-sectional population at τ were fixed throughout all simulations at 10% and 1 000 000, respectively. Estimators of the cumulative incidence were initiated at $t = 10$, that is, $\Lambda(t; 10) = \int_{10}^t \lambda(u) du$ was estimated, with onsets before $t = 10$ trimmed as discussed in § 2.4.

To illustrate the behaviour of $\hat{\Lambda}(t; 10)$, sample paths from 50 simulated datasets from scenario III are shown in Fig. 1 for effective sample sizes 250 and 1000. The effective sample size is the expected number $E(n_d)$ of prevalent cases recruited; for example, sampling 10 000 individuals yields an effective sample size of 1000 for prevalence proportion 10%. In Table 1, for each scenario considered, the empirical relative bias of $\hat{\Lambda}(t; 10)$ at $t = 20, 60, 100$ is provided for effective sample sizes 250, 500 and 1000. Similarly, the empirical coefficient of variation of $\hat{\Lambda}(t; 10)$ is provided in Table 2. For each sample size considered, 5000 datasets were generated to estimate relative biases and coefficients of variation. Figure 1 and Table 2 highlight the greater relative variability of the cumulative incidence at earlier times. In the Supplementary Material, the cumulative incidence is shown to exhibit, as expected, greater absolute variability at later times. The variability at all times decreases with sample size. Table 2 shows in most cases a very slight positive bias in small samples, which gradually vanishes in larger samples. Once more, the relative bias is greater at earlier times, while the absolute bias is greater at later times. In both

Fig. 1. Overlay of 50 sample paths of $\hat{\Lambda}(t; 10)$ for effective sample sizes 250 (a) and 1000 (b) under scenario III. The true curve is dashed and the mean estimated curve is solid.

Table 1. *The estimated percent relative bias of $\hat{\Lambda}(t; 10)$ at $t = 20, 60, 100$ and for effective sample size $E(n_d) = 250, 500, 1000$ under scenarios I, II, III and IV, using 5000 random datasets*

t	20	20	20	60	60	60	100	100	100
$E(n_d)$	250	500	1000	250	500	1000	250	500	1000
Scenario									
I	1.30	0.52	0.25	0.22	0.30	0.04	0.04	0.17	0.05
II	1.70	1.45	0.60	0.17	0.11	0.08	-0.01	0.06	-0.01
III	4.98	2.77	1.29	2.81	1.60	0.75	2.02	1.21	0.57
IV	3.78	2.51	1.57	2.21	1.05	0.51	1.22	0.70	0.31

Table 2. *The estimated percent coefficient of variation of $\hat{\Lambda}(t; 10)$ at $t = 20, 60, 100$ and for effective sample size $E(n_d) = 250, 500, 1000$ under scenarios I, II, III and IV, using 5000 random datasets*

t	20	20	20	60	60	60	100	100	100
$E(n_d)$	250	500	1000	250	500	1000	250	500	1000
Scenario									
I	39.7	27.7	19.9	10.9	7.7	5.5	7.02	5.0	3.6
II	53.9	37.5	26.4	11.7	8.2	5.8	6.6	4.6	3.3
III	53.5	35.3	24.9	24.3	16.8	11.2	18.7	13.4	8.6
IV	71.8	48.8	34.3	23.0	15.9	11.2	15.3	10.8	7.7

scenarios III and IV, the initial bias seems more pronounced. This observation may be a reflection of the fact that the rate of convergence is possibly slower under these scenarios due to the violation of the integral condition.

Quantile-quantile plots assessing the normality of the sampling distribution of $\hat{\Lambda}(t; 10)$ at $t = 30, 70$ and for effective sample sizes 250, 500 and 1000 were produced using 5000 random draws; these are included in the Supplementary Material. The sampling distribution generally

Table 3. *The estimated coverage probability (%) of bootstrapped 95% confidence intervals for $\Lambda(t; 10)$ at $t = 30, 70$ and for effective sample size $E(n_d) = 250, 500, 1000$ under scenarios I, II, III and IV, using 1000 random datasets and 1000 bootstrap repetitions for each dataset generated*

t	30	30	30	70	70	70
$E(n_d)$	250	500	1000	250	500	1000
Scenario						
I	92.3	95.2	94.4	93.2	94.8	94.9
II	91.7	93.8	94.7	93.4	94.5	95.2
III	94.1	94.9	94.1	94.2	93.3	94.9
IV	93.4	94.2	95.6	94.2	95.1	94.8

All the estimated standard errors associated with these coverage probability estimates were not greater than 1.79.

tends to be right-skewed, although this discrepancy decreases with sample size under scenarios I and II. Under these scenarios, even for smaller sample sizes, deviations from normality are rather mild and are more pronounced for $t = 30$. Under scenarios III and IV, however, deviations are more severe and do not decrease noticeably with sample size, a clear indication of the necessity of the integral condition in Theorem 1. A simulation study verifying the appropriateness of the bootstrap scheme proposed in § 3 reveals that the bootstrap percentile-based confidence intervals generally have very good coverage, as reported in Table 3. Even in cases with overt departure from normality, as under scenarios III and IV, only very minor under-coverage is observed, highlighting the robustness of the bootstrap scheme in this problem.

5. THE CANADIAN STUDY OF HEALTH AND AGING

In 1989, researchers from the University of Ottawa and Health Canada designed the Canadian Study of Health and Aging, a nationwide multicentre longitudinal study investigating the epidemiology of dementia in Canada. They were motivated by the impact of shifts in population age distributions on health services utilization. The researchers wished to determine the prevalence, incidence and risk factors of dementia, including Alzheimer's disease, in several subpopulations of Canada.

The study design included three stages. The first stage took place in 1991 and was the primary recruitment phase for the study. A total of 10 263 individuals of age 65 or higher were sampled from more than 36 communities, both rural and urban, across Canada, with simple random samples drawn from the community and a collection of institutions for the elderly. All provinces of Canada were represented. These individuals were assessed for demential conditions by the staff of the 18 participating field centres and followed up according to specific guidelines. The second and third stages of the study consisted primarily of reassessment, at 5- and 10-year marks, of individuals recruited earlier. Further design details are provided in McDowell et al. (1994).

Our focus is on using first stage data to infer the incidence of dementia in the Canadian elderly population. Using the methodology developed in this paper, we have investigated general trends in the overall incidence rate and standardized incidence rate of dementia, as well as trends among various subgroups. Linear extrapolation of the cumulative incidence was employed to estimate standardized incidence rates. We present the results pertaining to gender- and age-specific disease. Owing to the design-induced oversampling of the institutionalized elderly, correction for stratification was implemented, as discussed in § 3.3.

Fig. 2. Cumulative incidence of dementia in the Canadian elderly population, with bootstrapped 95% pointwise confidence intervals (dotted lines). The cumulative incidence for women (dash-dotted line) and men (dashed line) are superimposed.

Fig. 3. Cumulative age-specific standardized incidence of dementia in the Canadian elderly population, for age categories 65–70 (solid), 70–75 (dash), 75–80 (dot), 80–85 (dash-dot) and 85+ (long dash), from September 1976, 1979, 1981, 1983 and 1987 onwards, respectively.

Figure 2 presents the population-averaged cumulative incidence starting from September 1976, that is, the estimated number of disease onsets that have occurred since September 1976 in the elderly population as well as in the subpopulations of elderly men and women. Approximately 95% pointwise confidence intervals constructed using bootstrap resampling are provided for the overall cumulative incidence. This plot suggests that prior to 1981 an average of 92 000 new cases of dementia arose each year, roughly 49 000 of which were women, while after 1981 the average annual incidence rate fell to 70 000, with around 44 000 cases in women alone. It is unclear whether 1981 constituted a true pivotal year in the incidence of dementia or whether this change is an artefact of the inherent uncertainty of the onset recall process. Using the same data used here, [Addona & Wolfson \(2006\)](#) have tested the hypothesis of a constant incidence rate of dementia in the Canadian elderly population between 1976 and 1991 and concluded that the data did not provide evidence of a significant departure from this hypothesis. Figure 3 provides age-specific plots of the cumulative standardized incidence. The slope of each curve represents the number of

new cases per 1000 person-years at risk. The identifiability period in older age groups is significantly shorter than in younger age groups. A clear monotone ordering in age-specific standardized incidence rates emerges, with higher age groups experiencing a greater risk of dementia. The average age-specific standardized incidence rates were estimated to be 5.3, 15.6, 26.6, 48.2 and 152.6 cases per 1000 person-years for age categories 65–70, 70–75, 75–80, 80–85 and 85+, respectively. Previously reported rates estimated from incident cases in the second phase of the study, namely 5.5, 10.9, 22.3, 42.0 and 106.5 cases per 1000 person-years, respectively, were similar to our point estimates (CSHA, 2000), although our estimates and those in CSHA (2000) do not correspond to the same calendar period; observed discrepancies may be due to secular changes in incidence rates. While specific features found in Fig. 3 may provide interesting information about trends in age-specific standardized incidence rates, caution must be exercised when interpreting the local behaviour of the estimated curves. Indeed, large sudden increases in cumulative incidence curves are more likely to reflect that, in recalling onset, the next of kin tended to approximate distant onset dates using much less precise units of time. This observation does not hinder the interpretation of the behaviour of the estimated curves over larger intervals. This phenomenon is a limitation of this particular study and would not exist had onset information been obtained from medical records or other objective sources.

ACKNOWLEDGEMENT

The authors thank the editor and referees for their insightful and constructive comments, Professors Daniel O. Scharfstein, Thomas A. Louis and Christina Wolfson for helpful discussions, and Bin Zhu for providing the data. This work was supported by the National Science and Engineering Research Council of Canada, the Fonds Québécois de la Recherche sur la Nature et les Technologies, the National Institutes of Health and the Hopkins Sommer Scholars Program.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes additional technical details and simulation studies.

APPENDIX A

Relationship between left-truncation and onset process rate function

The process $N^*(t) = N\{\Lambda^{-1}(t)\}$ is a stationary Poisson process with unit rate for $t \geq \tau - b$. Using Ross (1983, Theorem 2.3.1) and the probability integral transform, given the occurrence of m events in $(\tau - b, \tau)$, the occurrence times $R_1, \dots, R_m$ of the process $\{N(t), t \geq \tau - b\}$ on $(\tau - b, \tau)$ are independent with the distribution function $\Lambda(r)/\Lambda(\tau)$. The truncation times $T_1^0 = \tau - R_1, \dots, T_m^0 = \tau - R_m$ are therefore independent with the density function $g(t) = \lambda(\tau - t)/\Lambda(\tau)$. Integrating this expression from $\tau - t$ to b , we find that $1 - G(\tau - t) = \kappa^{-1}\Lambda(t)$, where we have set $\kappa = \Lambda(\tau)$ and used that $G(b) = 1$ and $\Lambda(\tau - b) = 0$.

APPENDIX B

Uniform convergence of $\{\hat{\Lambda}(t; c) - \Lambda(t; c)\}/n_{\text{pop}}$ to zero in probability

It suffices to prove that $\{\hat{\Lambda}(t) - \Lambda(t)\}/n_{\text{pop}}$ is uniformly consistent. We define $\beta = \int S(u) dG(u)$ and set $\hat{\beta} = \int \hat{S}(u) d\hat{G}(u)$, where from Wang (1991, p. 133) we have that $\hat{G}(u) = \int_0^u \hat{S}(u)^{-1} d\hat{G}'(u)/$

$\int_0^\infty \hat{S}(u)^{-1} d\hat{G}'(u)$. In view of (3), we may write, for each $t \in [\tau - b, \tau]$,

$$\begin{aligned} \frac{\hat{\Lambda}(t) - \Lambda(t)}{n_{\text{pop}}} &= \frac{1 - \hat{G}(\tau - t)}{\hat{\beta}} \left\{ \frac{\hat{\Phi}(\tau) - \Phi(\tau)}{n_{\text{pop}}} \right\} + \frac{\Phi(\tau)}{n_{\text{pop}}} \{1 - \hat{G}(\tau - t)\} \left(\frac{1}{\hat{\beta}} - \frac{1}{\beta} \right) \\ &\quad - \frac{\Phi(\tau)}{\beta n_{\text{pop}}} \{\hat{G}(\tau - t) - G(\tau - t)\}, \end{aligned}$$

from which we find the inequality

$$\sup_{t \in [\tau - b, \tau]} \left| \frac{\hat{\Lambda}(t) - \Lambda(t)}{n_{\text{pop}}} \right| \leq \hat{\beta}^{-1} \left| \frac{\hat{\Phi}(\tau) - \Phi(\tau)}{n_{\text{pop}}} \right| + \frac{\Phi(\tau)}{n_{\text{pop}}} \left| \frac{1}{\hat{\beta}} - \frac{1}{\beta} \right| + \frac{\Phi(\tau)}{\beta n_{\text{pop}}} \sup_{u \in [0, b]} |\hat{G}(u) - G(u)|.$$

Both $\Phi(\tau)/n_{\text{pop}} = n\omega\phi(\tau)/n_{\text{pop}}$ and $\hat{\Phi}(\tau)/n_{\text{pop}} = \hat{\phi}(\tau)$ converge to $\phi(\tau)$ in probability as n_s tends to ∞ by the weak law of large numbers (van der Vaart, 2000, Example 2.1). The result follows from this last observation, the consistency of $\hat{\beta}$ (Wang, 1991, Lemma 4.2), the continuous mapping theorem (van der Vaart, 2000, Theorem 2.3) and the uniform consistency of $\hat{G}(u)$, proved in the Supplementary Material.

APPENDIX C

Weak convergence of $\hat{A}(t; c)$

The process $\{\hat{\Lambda}(t; c) - \Lambda(t; c)\}/n_{\text{pop}}$ admits the asymptotic representation

$$\begin{aligned} &\{\hat{\phi}(\tau) - \phi(\tau)\} \int_{\tau-t}^{\tau-c} \frac{dG'(u)}{S(u)} - \phi(\tau) \int_{\tau-t}^{\tau-c} \{\hat{S}(u) - S(u)\} \frac{dG'(u)}{S(u)^2} \\ &\quad + \phi(\tau) \int_{\tau-t}^{\tau-c} \frac{1}{S(u)} d\{\hat{G}'(u) - G'(u)\} + r_{n_s}(t), \end{aligned}$$

where, as proved in the Supplementary Material, $\lim_{n_s, n} n_s^{1/2} \sup_{t \in [c, \tau]} |r_{n_s}(t)| = 0$ in probability. This representation is key to proving, as in the Supplementary Material, that $\hat{A}(t; c)$ converges weakly to a mean-zero process $A(t; c)$ with the same law as

$$B_0 \int_{\tau-t}^{\tau-c} \frac{dG'(u)}{S(u)} - \{\phi(\tau)\}^{1/2} \left[\int_{\tau-t}^{\tau-c} B_1(u) \frac{dG'(u)}{S(u)^2} - \int_{\tau-t}^{\tau-c} \frac{1}{S(u)} dB_2(u) \right], \quad (\text{A1})$$

where B_0 is a mean zero Gaussian variate with variance $\phi(\tau)\{1 - \phi(\tau)\}$, B_1 is a mean zero Gaussian process with the covariance function

$$\Psi(u, v) = S(u)S(v) \int_0^{u \wedge v} \frac{dF(x)}{R(x)S(x)},$$

where $F(x) = 1 - S(x)$ and $R(x) = \text{pr}(T \leq x \leq Y)$ (Woodroffe, 1985, pp. 174–6; Tsai et al., 1987; Wang, 1991, p. 135), and B_2 has the law of $B_3 \circ G'$, where B_3 is a Brownian bridge. As in Addona et al. (2009, Theorem 2), B_0 is orthogonal to (B_1, B_2) , while in view of Wang (1991, Lemma 4.5), B_1 and B_2 are orthogonal. Consequently, the covariance function $\Sigma(s, t)$ of $A(t; c)$ can be decomposed as $\Sigma_0(s, t) + \phi(\tau)\{\Sigma_1(s, t) + \Sigma_2(s, t)\}$, each obtained from one of the three summands in (A1). As derived in the Supplementary Material, we find that

$$\begin{aligned} \Sigma_0(s, t) &= \phi(\tau)\{1 - \phi(\tau)\}\{G(\tau - c) - G(\tau - s)\}\{G(\tau - c) - G(\tau - t)\}/\beta^2, \\ \Sigma_1(s, t) &= \int_0^{\tau-c} \frac{[G(\tau - c) - G\{(\tau - s) \vee x\}][G(\tau - c) - G\{(\tau - t) \vee x\}]}{R(x)S(x)} dF(x)/\beta^2, \end{aligned}$$

$$\Sigma_2(s, t) = \left[\beta \int_{\tau-s}^{\tau-c} \frac{dG(u)}{S(u)} - \{G(\tau-c) - G(\tau-s)\} \{G(\tau-c) - G(\tau-t)\} \right] / \beta^2$$

for $s \leq t$. Estimation of all the quantities involved in $\Sigma(s, t)$ has been discussed so far, aside from $F(x)$ and $R(x)$, for which $\hat{F}(x) = 1 - \hat{S}(x)$ and $\hat{R}(x) = n_d^{-1} \sum_{i=1}^{n_d} \mathbb{I}_{[T_i, Y_i]}(x)$ are uniformly consistent estimators, respectively.

REFERENCES

- ADDONA, V. & WOLFSON, D. B. (2006). A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up. *Lifetime Data Anal.* **12**, 267–84.
- ADDONA, V., ASGHARIAN, M. & WOLFSON, D. B. (2009). On the incidence-prevalence relation and length-biased sampling. *Can. J. Statist.* **37**, 206–18.
- ALHO, J. M. (1992). On prevalence, incidence, and duration in general stable populations. *Biometrics* **48**, 587–92.
- ALIOUM, A., COMMENGES, D., THIÉBAUT, R. & DABIS, F. (2005). A multistate approach for estimating the incidence of human immunodeficiency virus by using data from a prevalent cohort study. *Appl. Statist.* **54**, 739–52.
- ASGHARIAN, M., M'LAN, C. E. & WOLFSON, D. B. (2002). Length-biased sampling with right-censoring: an unconditional approach. *J. Am. Statist. Assoc.* **97**, 201–10.
- BROOKMEYER, R. & DAMIANO, A. (1989). Statistical methods for short-term projections of AIDS incidence. *Statist. Med.* **8**, 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. *Am. J. Epidemiol.* **141**, 166–72.
- CHENG, M. Y., HALL, P. & YANG, Y. J. (2007). Nonparametric inference under dependent truncation. *Acta Sci. Math.* **73**, 397–422.
- COX, D. R. & OAKES, D. (1984). *Analysis of Survival Data*. Boca Raton: Chapman & Hall/CRC.
- CSHA (2000). The incidence of dementia in Canada. *Neurology* **55**, 66–73.
- ESBJERG, S., KEIDING, N. & KOCH-HENRIKSEN, N. (1999). Reporting delay and corrected incidence of multiple sclerosis. *Statist. Med.* **18**, 1691–706.
- FREEMAN, J. & HUTCHISON, G. B. (1980). Prevalence, incidence and duration. *Am. J. Epidemiol.* **112**, 707–23.
- KAPLAN, E. L. & MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Am. Statist. Assoc.* **53**, 457–81.
- KEIDING, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *J. R. Statist. Soc. A* **154**, 371–412.
- KEIDING, N., HOLST, C. & GREEN, A. (1989). Retrospective estimation of diabetes incidence from information in a prevalent population and historical mortality. *Am. J. Epidemiol.* **130**, 588–600.
- 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. *Can. Med. Assoc. J.* **150**, 899–912.
- MIETTINEN, O. (1976). Estimability and estimation in case-referent studies. *Am. J. Epidemiol.* **103**, 226–35.
- ROSS, S. M. (1983). *Stochastic Processes*. New York: John Wiley & Sons.
- ROTHMAN, K. J., GREENLAND, S. & LASH, T. L. (2008). *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins.
- TSAI, W., JEWELL, N. P. & WANG, M. C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **74**, 883–6.
- VAN DER VAART, A. W. (2000). *Asymptotic Statistics*. New York: Cambridge University Press.
- VAN DER VAART, A. W. & WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.
- WANG, M. C. (1991). Nonparametric estimation from cross-sectional survival data. *J. Am. Statist. Assoc.* **86**, 130–43.
- WANG, M. C., BROOKMEYER, R. & JEWELL, N. P. (1993). Statistical models for prevalent cohort data. *Biometrics* **49**, 1–11.
- WOLFSON, C., WOLFSON, D. B., ASGHARIAN, M., M'LAN, C. E., OSTBYE, T., ROCKWOOD, K. & HOGAN, D. B., FOR THE CLINICAL PROGRESSION OF DEMENTIA STUDY GROUP (2001). A reevaluation of the duration of survival after the onset of dementia. *New Engl. J. Med.* **344**, 1111.
- WOODROOFE, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13**, 163–77.

[Received March 2009. Revised March 2012]