A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia.

Pierre Joly, Daniel Commenges, Catherine Helmer, Luc Letenneur

A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia

Pierre Joly, Daniel Commenges, Catherine Helmer, Luc Letenneur

ISPED

Université de Bordeaux II

146, rue Léo Saignat

33076 Bordeaux Cedex, France

Tel : (33) 5 57 57 45 16

Fax : (33) 5 56 24 00 81

E-mail: Pierre.Joly@isped.u-bordeaux2.fr

Corresponding author: Pierre Joly

URL: www.isped.u-bordeaux2.fr/isped/recherche/biostats/fr-biostats-accueil.htm

December 13, 2001

Summary

We consider the problem of estimating the intensity functions for a continuous time "illness-death" model with intermittently observed data. In such a case, it may happen that a subject becomes diseased between two visits and dies without being observed. Consequently, there is an uncertainty about the precise number of transitions. Estimating the intensity of transition from health to illness by survival analysis (treating death as censoring) is biased downwards. Furthermore, the dates of transitions between states are not known exactly. We propose to estimate the intensity functions by maximising a penalized likelihood. The method yields smooth estimates without parametric assumptions. This is illustrated using data from a large cohort study on cerebral ageing. The age-specific incidence of dementia is estimated using an illness-death approach and a survival approach.

Key words: Illness-death model, intensity function, interval-censoring, penalized likelihood, splines, truncation.

1 Introduction

The three-state "illness-death" model is useful in a variety of biomedical settings to characterize the risk of an individual, especially when competing risks between death and a disease are assumed (Kodell and Nelson, 1980; McKnight and Crowley, 1984; Andersen, 1988; Lindsey and Ryan, 1993; Andersen et al., 1993). Data analysis in this context is complicated by the fact that subjects are observed only intermittently: a subject healthy at a visit and deceased before the next scheduled visit may have transited to the "illness" state without being diagnosed. On the one hand, the dates of transition between states may not be known exactly and on the other, the number of transitions between states may not be known exactly. As explained in section 2, this situation may lead to an underestimation of the incidence of a disease. The non-parametric approach developed by Frydman (1995) considers only the particular case where the status of the subject is known before his death; in addition the non-parametric maximum likelihood approach does not yield directly estimators of the transition intensities. In this paper, we propose a non-parametric penalized likelihood method for estimating the intensities in an "illness-death" model for intermittently observed data. Intensity functions yield a very interesting description of epidemiologic processes since they may often be interpreted as incidence or mortality rates (Keiding, 1991; Hougaard, 1999).

The epidemiological motivations for this work are described in section 2. Then we describe how to estimate the incidence. We introduce smoothness assumptions on the intensity functions by penalizing the log-likelihood with a term which has large values for rough functions. The estimators are defined non-parametrically as the functions which maximize the penalized likelihood. As the maximum penalized likelihood estimators cannot be calculated explicitly, they are approximated using splines. This approach, presented in section 3, is an extension to an illness-death model of survival analysis approaches, proposed by O'Sullivan (1988), Gu (1996), and Joly et al. (1998). The models proposed are a gener-

alization of the progressive three-state model proposed by Joly and Commenges (1999). In section 4, we present a simulation study. In section 5, the proposed model is applied to the estimation of the age-specific incidence of dementia based on data from a large prospective cohort study.

2 Epidemiological motivations

In a cohort study such as that presented in section 5 where subjects are only intermittently observed, they may develop the disease and die between two visits. The apparent transition is from "health" to "death" while two transitions have in fact occurred: "health" to "illness" then "illness" to "death". This leads to an uncertainty about the precise number of diseased subjects at risk of death. For example, in figure 1, we present the follow-up of a subject in a cohort study: V_0, V_1, ..., V_m are times of follow-up. We suppose that the subject develops the disease after V_m and then dies before V_{m+1}.

Information about the subject are the date of entry in the study (V_0), the date (V_m) of the last visit seen in the healthy state and the date of death (T). However, whether he is ill or healthy at the time of his death is unknown. When the incidence of a disease is studied using a two-state survival analysis, the subject is not observed as "ill" and his survival time is considered as right-censored at V_m. This leads to an underestimation of the incidence of the disease.

To understand the nature of the bias, we consider a very simple parametric model with three constant intensity functions (exponential law): α_{01}, α_{02} and α_{12} (we consider here a particular case of the model depicted in figure 2). We consider that observations are taken at only two times, 0 and t_1. When attempting to estimate α_{01} using a (two-state) survival analysis, subjects who died between 0 and t_1 are considered as right-censored at 0. With independent censoring one

consistent estimator of the survival function at t_1 would be:

$$\frac{\#\text{healthy at } t_1}{\#\text{alive at } t_1}$$

This estimator tends (as the sample size $n \to \infty$) towards:

$$\tilde{S}(t_1) = \frac{\mathrm{P}(\text{healthy at } t_1)}{\mathrm{P}(\text{alive at } t_1)} \quad .$$

Any other consistent estimator in this two-state survival model would tend to $\tilde{S}(t_1)$. Because in the exponential model $S(t_1) = e^{-\alpha_{01} t_1}$, estimators of α_{01} by this method will tend towards: $\tilde{\alpha}_{01} = -t_1 log(\tilde{S}(t_1))$. On the other hand we can compute with the true "illness-death" model:

$$\frac{\mathrm{P}(\text{healthy at } t_1)}{\mathrm{P}(\text{alive at } t_1)} = \frac{1}{1 + \frac{\alpha_{01}}{\alpha_{01}+\alpha_{02}-\alpha_{12}}(e^{(\alpha_{01}+\alpha_{02}-\alpha_{12})t_1} - 1)} \quad .$$

Hence we deduce the asymptotic bias of estimators based on the two-state survival model:

$$\tilde{\alpha}_{01} - \alpha_{01} \quad = \quad \frac{1}{t_1} log \left(\frac{\alpha_{01} e^{-(\alpha_{12}-\alpha_{02})t_1} - (\alpha_{12} - \alpha_{02}) e^{-\alpha_{01} t_1}}{\alpha_{01} - (\alpha_{12} - \alpha_{02})} \right) . \qquad (1)$$

If both mortality rates α_{02} and α_{12} are null, the bias is null, which is obvious since the illness-death model then reduces to a (two-state) survival model. In the general case we see that a key parameter is the differential mortality $\Delta = \alpha_{12} - \alpha_{02}$; the bias is null if this differential mortality is null. In general Δ is positive (because ill people have a higher mortality rate) and it can be shown that the bias is negative and that it increases with t_1 and Δ. This formula can be used to obtain an idea of the bias for given values of the intensity functions and the length of the intervals.

In this paper we propose a penalized likelihood approach for estimating the intensities in an "illness-death" model for intermittently observed data in a more general context. This approach allows us to: i) use the full likelihood of the problem; ii) obtain smooth estimates of the intensity functions; iii) be free of parametric assumptions.

3 The Model

3.1 The illness-death model

The three-state model with irreversible transitions is shown in Figure 2. State 0 represents the state "Health", state 1 "Illness" and state 2 "Death".

We define α_{ij} as the intensity functions and A_{ij} as the corresponding cumulative intensity functions, $A_{ij}(t) = \int_0^t \alpha_{ij}(u)du$.

The intensities α_{01} and α_{02} may depend on t (the age or the calendar time) and the intensity α_{12} may depend on t and on τ (the time since the onset of the disease). As we are mostly interested in age-specific incidence and age-specific mortality, in this paper, we assume for simplicity that the intensities only depend on age, so we use a non-homogeneous Markov model. The transition intensity α_{01} represents the age-specific incidence of illness whilst the transition intensities α_{12} and α_{02} represent the age-specific mortality rates for ill and healthy subjects respectively.

3.2 The observations

Vital status and time of death are known exactly. However, at T_i which is the age of death or the end of the study, the disease status ("healthy" or "ill") may be unknown. Subject i is seen at V_{ik}, $k = 0, ..., m_i$, $m_i \geq 0$ and $V_{im_i} \leq T_i$; V_{i0} is called the age at the baseline visit. The disease status is assessed only at these visits. Thus if a transition towards illness is observed, the age at the time of transition is interval-censored; if a subject who dies was healthy at the last visit, it is not known whether he has made the transition towards illness or not. Since the subjects are not observed from birth but rather from the beginning of the study, the time spent in state 0 is generally left-truncated, because subjects must be in the "health" state at the begining of the study to be included in the sample. We assume that the truncating and censoring mechanisms are independent from

the illness-death process. This happens for instance if $V_{i0}, V_{i1}, ..., V_{im_i}$ are fixed or are random variables independent of the illness-death process, and if loss of follow-up obeys an independent censoring mechanism.

3.3 The likelihood

In the application, the four possible different cases for the observation of a subject are: i) observed healthy during the first visits of the follow-up and still alive at the end of the study (at this date we do not know if the subject is ill or healthy); ii) observed healthy and is then deceased (we do not know if the subject was ill or healthy at the time of his death); iii) observed ill at one follow-up time and is still alive at the end of the study; iv) observed ill at one follow-up time and then is deceased.

The likelihood contributions to the four different cases of the observation are detailed in Appendix A.

3.4 The penalized likelihood

Intensity functions are expected to be smooth. To introduce such *a priori* knowledge, we penalize the likelihood by a term which has large values for rough functions. The roughness penalty function chosen for the three-state model is the sum of the square norms of the second derivatives of the intensities. The penalized log-likelihood (pl) is thus defined as

$$pl(\alpha_{01}, \alpha_{12}, \alpha_{02}) = l(\alpha_{01}, \alpha_{02}, \alpha_{12}) - \kappa_{01}\int \alpha_{01}^{''\,2}(u)du - \kappa_{12}\int \alpha_{12}^{''\,2}(u)du - \kappa_{02}\int \alpha_{02}^{''\,2}(u)du \quad (2)$$

where l is the full log-likelihood (which is a function of $\alpha_{01}(.)$, $\alpha_{12}(.)$ and $\alpha_{02}(.)$) and κ_{01}, κ_{12} and κ_{02} are three positive smoothing parameters which control the trade-off between the data fit and the smoothness of the functions. Maximization of (2) defines the maximum penalized likelihood estimators (MPLE) $\hat{\alpha}_{01}(.)$, $\hat{\alpha}_{12}(.)$ and $\hat{\alpha}_{02}(.)$.

3.5 Approximation of the estimators

The MPLE cannot be calculated explicitly. However, it can be approximated using splines. Splines are piecewise polynomial functions which are combined linearly to approximate a function on an interval. We use cubic M-splines and I-splines, which are variants of B-splines. For more details, see Joly et al. (1998).

The estimator $\hat{A}(.)$ for a given transition is approximated by a linear combination of m I-splines: $\tilde{A}(.) = \tilde{\boldsymbol{\theta}}\boldsymbol{I}(.)$, where $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, ..., \tilde{\theta}_m)$ and $\boldsymbol{I}(.) = (I_1(.), ..., I_m(.))^T$. By differentiation we obtain: $\tilde{\alpha}(.) = \tilde{\boldsymbol{\theta}}\boldsymbol{M}(.)$, where $\boldsymbol{M}(.) = (M_1(.), ..., M_m(.))^T$. We use a distinct base of splines for each intensity function, possibly with a different number of splines in each basis. The monotonicity constraint for $\tilde{A}(.)$ is fulfilled by constraining the coefficients $\tilde{\boldsymbol{\theta}}$ to be positive. The approximation $\tilde{\alpha}$ of $\hat{\alpha}$ is the function belonging to the space generated by the basis of splines, which maximizes $pl(\alpha_{01}, \alpha_{12}, \alpha_{02})$. The vectors of spline coefficients $\tilde{\boldsymbol{\theta}}_{01}$, $\tilde{\boldsymbol{\theta}}_{12}$ and $\tilde{\boldsymbol{\theta}}_{02}$ for fixed κ_{01}, κ_{12} and κ_{02} are obtained simultaneously by maximizing the log-likelihood using Marquardt's algorithm (1963). When the three vectors $\tilde{\boldsymbol{\theta}}_{01}$, $\tilde{\boldsymbol{\theta}}_{12}$ and $\tilde{\boldsymbol{\theta}}_{02}$ are obtained, all the functions of interest can be computed, as in a parametric method.

3.6 Confidence intervals

A Bayesian technique for generating confidence bands for penalized likelihood estimators was proposed by O'Sullivan (1988) for survival analysis. Therefore, $\boldsymbol{\theta_{ij}}$, $(i,j) = (0,1), (1,2), (0,2)$ is regarded as a random variable. Up to a constant, the penalized log-likelihood pl is a posterior log-likelihood for $\boldsymbol{\theta_{ij}}$ and the penalty term is the prior log-likelihood. After a Gaussian approximation, the covariance of $\boldsymbol{\theta_{ij}}$ is $-\hat{\boldsymbol{H}}_{\boldsymbol{(ij)}}^{-1}$, where $\hat{\boldsymbol{H}}_{\boldsymbol{(ij)}}$ is the converged Hessian $\frac{\partial^2 pl}{\partial \boldsymbol{\theta_{ij}}^2}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{ij}})$. Therefore, an approximate 95% Bayesian confidence interval for $\tilde{\alpha}_k$ at point t is:

$$\tilde{\alpha}_{ij}(t) \pm 1,96\sqrt{\boldsymbol{M_{(ij)}}(t)^T \left[-\hat{\boldsymbol{H}}_{\boldsymbol{(ij)}}^{-1}\right] \boldsymbol{M_{(ij)}}(t)},$$

For the three-state model proposed, we simply used this method for each of the three transitions. However, this estimator does not take into account the variability due to the choice of smoothing parameters.

3.7 Selection of smoothing parameters

O'Sullivan (1988) proposed an approximate cross-validation score for survival models. We extend his method to our case to choose the three smoothing parameters simultaneously.

The standard cross-validation score which must be maximized to obtain $\boldsymbol{\kappa} = (\kappa_{01}, \kappa_{12}, \kappa_{02})$ is:

$$V(\boldsymbol{\kappa}) = \sum_{i=1}^{n} l_i(\tilde{\boldsymbol{\theta}}_{-i})$$

where $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\theta}}_{01}, \tilde{\boldsymbol{\theta}}_{12}, \tilde{\boldsymbol{\theta}}_{02})$ and $\tilde{\boldsymbol{\theta}}_{-i} = \tilde{\boldsymbol{\theta}}_{-i}(\boldsymbol{\kappa})$ is the maximum penalized likelihood estimator of $\boldsymbol{\theta}$ for the sample in which the i^{th} individual is removed and l_i is the log-likelihood contribution of this individual. The maximization of this score is computationally prohibitive, however, since it requires a maximization for each individual and for each different value of $\boldsymbol{\kappa}$. So we maximize the approximate score:

$$\overline{V}(\boldsymbol{\kappa}) \quad = \quad l(\tilde{\boldsymbol{\theta}}) - trace\left(\hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{H}}_{\boldsymbol{l}}\right) \tag{3}$$

with $\hat{\boldsymbol{H}} = \dfrac{\partial^2 pl}{\partial \boldsymbol{\theta}^2}(\tilde{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{H}}_{\boldsymbol{l}} = \dfrac{\partial^2 l}{\partial \boldsymbol{\theta}^2}(\tilde{\boldsymbol{\theta}})$.

4 Simulation study

The aim of the simulation study was to evaluate our method and compare it to the (two-state) survival analysis approach for the estimation of α_{01}. The latter approach treats death as censoring and should lead to an underestimation of α_{01} (as studied in the simple example of section 2). The method was that proposed in

Joly et al. (1998) and Commenges et al. (1998): it also used penalized likelihood and treated interval-censoring (but not in the framework of the illness-death model).

The data were generated from a mixture of gamma distributions for α_{01}, and simple Weibull distributions for α_{12} and α_{02}. We generated the times of follow-up V_0, V_1, ..., V_m for each subject in order to simulate the interval-censoring process; $V_0 = 0$ and $V_{k+1} = V_k + 1 + 3u$, where u was uniformly distributed on $[0, 1]$. The end of the study was generated for each subject as $2 + 50u$.

The density corresponding to α_{01} was a Gamma mixture $(0.4\Gamma(t; 37, 1.5) + 0.6\Gamma(t; 20, 2))$, with the probability density functions $\Gamma(t; \alpha, \gamma) = \frac{\alpha^\gamma t^{\gamma-1} e^{-\alpha t}}{\Gamma(\gamma)}$. The densities corresponding to α_{12} and α_{02} were Weibull $(f(t; 2.5, 0.08))$ and $(f(t; 3, 0.04))$ respectively, with the probability density functions $f(t; \gamma, p) = p\gamma^p t^{p-1} e^{-(\gamma t)^p}$ (figure 3).

The sample consisted of 3000 subjects; 965 were classified as "ill" and 887 of them died. A total of 1181 subjects died without being diagnosed as "ill".

Due to interval-censoring, 358 subjects died coming from the illness state but without being classified as "ill" before. In the (two-state) survival analysis, they were considered as right-censored for α_{01} at the time of the last visit.

Smoothing parameters for the two models were chosen using a cross-validation method. We used 12 knots and cubic splines for the approximation of α_{01} with the two models. In the "illness-death model" we used 7 knots and cubic splines for the two other intensity functions. Figure 4 displays the estimation of α_{01} for one simulated example from a Gamma mixture. The solid line represents the true intensity function of α_{01}. The dashed line represents the estimate of α_{01} with the "illness-death model" and the dotted line represents the estimate of α_{01} using a (two-state) survival analysis. For each time, the latter is lower than the former. This exemplifies the underestimation of α_{01} using a (two-state) survival analysis approach in the case of interval-censoring and a high risk of death for ill subjects.

5 Application

The application is based on the Paquid research programme (Letenneur et al., 1999), a prospective cohort study of mental and physical aging that evaluates social environment and health status. The target population consists of subjects aged 65 years and older living at home in southwestern France. The baseline variables registered included socio-demographic factors, medical history and psychometric tests. Subjects were re-evaluated 1, 3, 5 and 8 years after the initial visit. Prevalent cases were removed from the sample because of their high mortality relative to non-demented people of the same age. Therefore, this produced a left-truncation problem. The sample consisted of 3672 subjects, of whom 1882 were between 65-75 years old and 1419 between 75-85, whilst 371 were 85 and older at the baseline visit. During the 8 years of follow-up, 281 incident cases of dementia were observed of whom 113 died; 1077 subjects were observed in the healthy state at the last visit before death. The basic time scale was age so the transition intensity α_{01} represents the age-specific incidence of dementia. The age of onset of dementia of the subjects was left-truncated by the age at inclusion in the study. The information available on incident cases of dementia was the date of the visit last seen without dementia and the date of the visit first seen with dementia. Thus, the age of onset of dementia was interval-censored. The dates of death were known exactly. The non-homogeneous Markov model discussed in this paper allows the direct treatment of these data, which are both interval-censored and left-truncated.

We compared two methods for estimating the age-specific incidence of dementia: the first one, which has been previously used (Commenges et al., 1998), was a (two-state) survival analysis dealing with interval-censored data, and the second was the illness-death model proposed here. We used 7 knots and cubic splines for the approximation of each intensity function. The approximate cross-validation method leads to $\kappa = 6.5\ 10^6$ (for the two-state survival model) and to

$\kappa_{01} = 3.5\,10^5$, $\kappa_{12} = 6.5\,10^3$ and $\kappa_{02} = 3.2\,10^5$. Figure 5 displays the estimated age-specific incidence of dementia. For each age, the risk of developing dementia estimated with the "illness-death model" is higher than the risk estimated using a (two-state) survival model.

Commenges et al. (1998) found that the age-specific incidence of dementia was higher in men than in women under 76 years old, and higher in women than in men when they were older. Since the age-specific risk of death is higher in men than in women, whatever their age, the difference between the age-specific incidence of dementia between men and women could be due to the differential underestimation; it might be more underestimated in men than in women, especially in old age. This hypothesis can be examined with our model. Figure 6 shows the age-specific incidence of dementia estimated separately for men and women by the "illness-death model". For each age, the risk of developing dementia evaluated with the "illness-death model" was higher than that estimated separately for men and for women using the (two-state) survival model of Commenges et al. (1998). However, the age-specific incidence of dementia in figure 6 was higher in men than in women under 80 years old, and higher in women than in men after this age. Therefore, it is unlikely that the higher incidence observed among older women is explained fully by the higher mortality among men.

The three intensity functions α_{01}, α_{12} and α_{02} were estimated simultaneously by the "illness-death model". Figure 7 displays all the intensities estimated. For each age, there is a higher risk of death than of developing dementia, and there is a higher risk of death for demented people.

6 Discussion

With interval-censoring and a high risk of death for diseased subjects, there is an underestimation of α_{01} using a (two-state) survival analysis approach, which must be taken into account. We have shown that the penalized likelihood approach in

an illness-death model yields a method for analyzing such data and for providing estimators of the intensity functions, which cannot be correctly estimated using conventional non-parametric methods. The proposed approach can be applied to semi-Markov models as well as to non-homogeneous Markov models.

As suggested in Joly and Commenges (1999), the penalized likelihood can be used to estimate the intensity functions in a regression model. The roughness penalty function is the sum of the square norms of the second derivatives of the baseline intensity functions. The regression parameters and the baseline functions are estimated simultaneously.

Here we have used a non-homogeneous Markov model. A more satisfactory model would be to assume that the transition intensity α_{12} depends on both age and duration of the disease. A direct approach is to model non-parametrically a function of these two different times; it is computationally difficult to estimate, although it has been attempted in some problems (Gu, 1996; Hansen et al., 1998). It would be interesting to develop more restrictive models for $\alpha_{12}(t, \tau)$ which are easier to estimate.

Whatever the model, the time between two observations must not be too large if one wants reliable estimates. The main parameter to assess the length of an interval is probably the risk of dying of deseased subjects during this interval.

REFERENCES

Andersen, P.K. (1988). Multistate models in survival analysis: a study of nephropathy and mortality in diabetes. *Statistics in Medicine* **7**, 661-670.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical models based on counting processes.* Springer-Verlag, New-York.

Commenges, D., Letenneur, L., Joly, P., Alioum, A. and Dartigues, J-F. (1998). Modelling age-specific risk: application to dementia. *Statistics in Medicine* **17**, 1973-1988.

Frydman, H. (1995). Nonparametric estimation of a Markov "illness-death" process from interval-censored observations, with application to diabetes survival data. *Biometrika* **82**, 773-789.

Gu, C. (1996). Penalized likelihood hazard estimation: a general procedure. *Statistica Sinica* **6**, 861-876.

Hansen, M., Kooperberg, C. and Sardy, S. (1998). Triogram Models. *Journal of the American Statistical Association* **93**, 441, 101-119.

Hougaard, P. (1999). Fundamentals of survival data. *Biometrics* **55**, 13-22.

Joly, P., Commenges, D. and Letenneur, L. (1998). A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics* **54**, 185-194.

Joly, P. and Commenges, D. (1999). A penalized likelihood approach for a progressive three-state model with censored and truncated data: application to AIDS. *Biometrics* **55**, 887-890.

Keiding, N. (1991). Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society, Series A* **154**, Part 3, 371-412.

Kodell, R. L., and Nelson, C. J. (1980). An illness-death model for the study of the carcinogenic process using survival/sacrifice data. *Biometrics* **36**, 267-277.

Letenneur, L., Gilleron, V., Commenges, D., Helmer, C., Orgogozo, J.-M. and Dartigues, J.-F. (1999). Are sex and educational level independent predictors of dementia and Alzheimer's disease ? Incidence data from the PAQUID project. *Journal of Neurology Neurosurgery and Psychiatry* **66**, 177-183.

Lindsey, J. C. and Ryan, L. M. (1993). A three-state multiplicative model for rodent tumorigenicity experiments. *Journal of the Royal Statistical Society, Series C* **42**, 283-300.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal of Applied Mathematics* **11**, 431-441.

McKnight, B. and Crowley, J. (1984). Tests for differences in tumor incidence based on animal carcinogenesis experiments. *Journal of the American Statistical Association* **79**, 639-648.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal on Scientific and Statistical Computing* **9**, 363-379.

Appendix A

For the sake of simplicity we omit the index i. The subject is observed at $V_0, V_1, ..., V_m$. T is the age at death or at the end of the study. The likelihood contribution of one subject for the four possible observed trajectories is:

i) subject enters the study at V_0 and is healthy. He is healthy at V_m and still alive at T (no information about illness at T):

$$L = \frac{1}{e^{-A_{01}(V_0)-A_{02}(V_0)}} \left\{ e^{-A_{01}(T)-A_{02}(T)} + \int_{V_m}^{T} e^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) e^{-(A_{12}(T)-A_{12}(u))} du \right\} .$$

In the application we often have $V_m = T$ and in this particular case:

$$L = \frac{e^{-A_{01}(V_m)-A_{02}(V_m)}}{e^{-A_{01}(V_0)-A_{02}(V_0)}} .$$

ii) subject enters the study at V_0 and is healthy. He is healthy at V_m and dies at T (we do not know if he is ill or healthy at the time of his death):

$$L = \frac{1}{e^{-A_{01}(V_0)-A_{02}(V_0)}} \left\{ e^{-A_{01}(T)-A_{02}(T)} \alpha_{02}(T) + \int_{V_m}^{T} e^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) e^{-(A_{12}(T)-A_{12}(u))} \alpha_{12}(T) du \right\} .$$

iii) subject enters the study at V_0 and is healthy. He is healthy at V_k, $(k < m)$, ill at V_{k+1}, and still alive at T:

$$L = \frac{1}{e^{-A_{01}(V_0)-A_{02}(V_0)}} \int_{V_k}^{V_{k+1}} e^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) e^{-(A_{12}(T)-A_{12}(u))} du .$$

iv) subject enters the study at V_0 and is healthy. He is healthy at V_k, $(k < m)$, ill at V_{k+1}, and dies at T:

$$L = \frac{1}{e^{-A_{01}(V_0)-A_{02}(V_0)}} \int_{V_k}^{V_{k+1}} e^{-A_{01}(u)-A_{02}(u)} \alpha_{01}(u) e^{-(A_{12}(T)-A_{12}(u))} \alpha_{12}(T) du .$$

Captions for Figures

Figure 1: follow-up of a subject

Figure 2: The illness-death model: α_{01}, α_{12} and α_{02} are the intensity functions.

Figure 3: True and estimated intensity functions α_{12} and α_{02}. The upper solid line represents the true intensity function α_{12}, the lower solid line represents the true intensity function α_{02}, the upper dashed line represents the estimated $\hat{\alpha}_{12}$ and the lower dashed line represents the estimated $\hat{\alpha}_{02}$.

Figure 4: True and estimated intensity functions α_{01}. The solid line represents the true intensity function α_{01}, the dashed line represents $\hat{\alpha}_{01}$ estimated with an illness-death model and the dotted line represents $\hat{\alpha}_{01}$ estimated using a (two-state) survival analysis.

Figure 5: Estimated age-specific incidence of dementia. The solid line represents the intensity function α_{01} estimated using the "illness-death model" and the dotted line represents the intensity function estimated using (two-state) a survival analysis. Paquid 1999

Figure 6: Estimated age-specific incidence of dementia (women = solid line and men = dashed line). Paquid 1999

Figure 7: Estimation of the age-specific incidence of dementia (solid line), of the intensity function of death for demented subjects (dashed line) and of the intensity function of death for non-demented subjects (dotted line). Paquid 1999

Figure 1

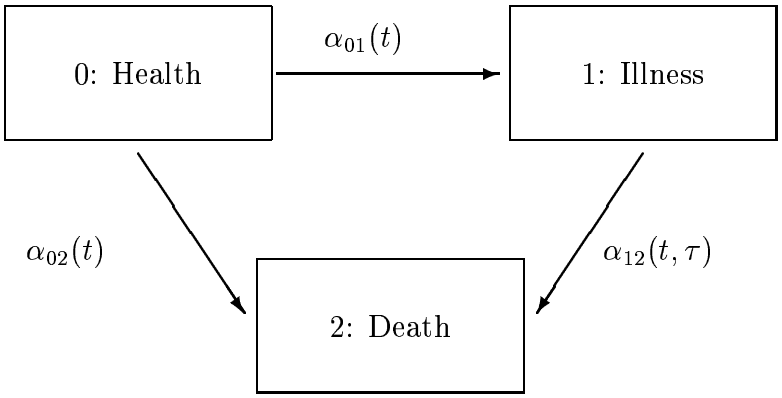

Figure 2

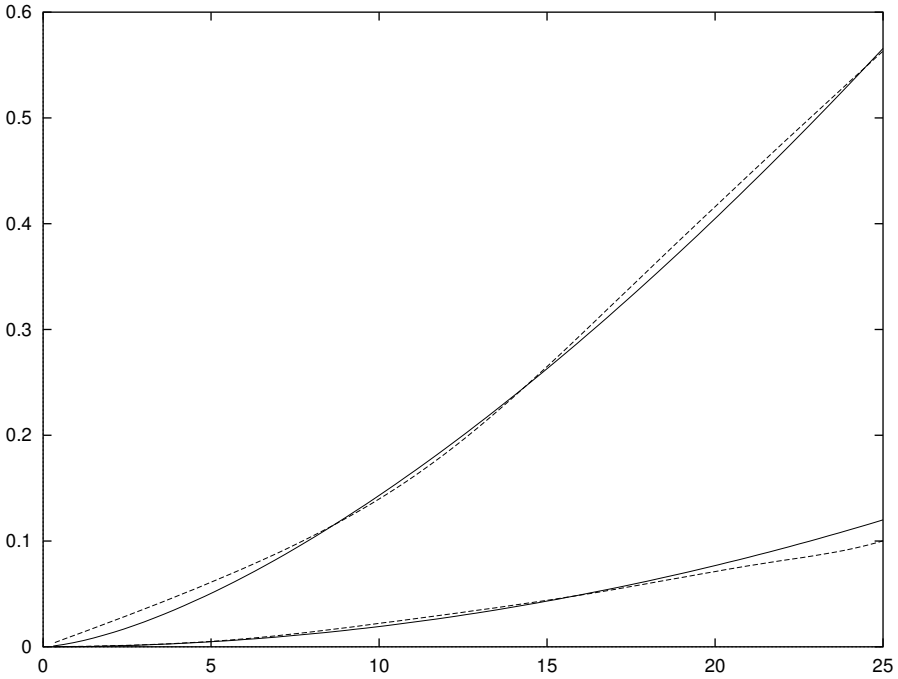

Figure 3

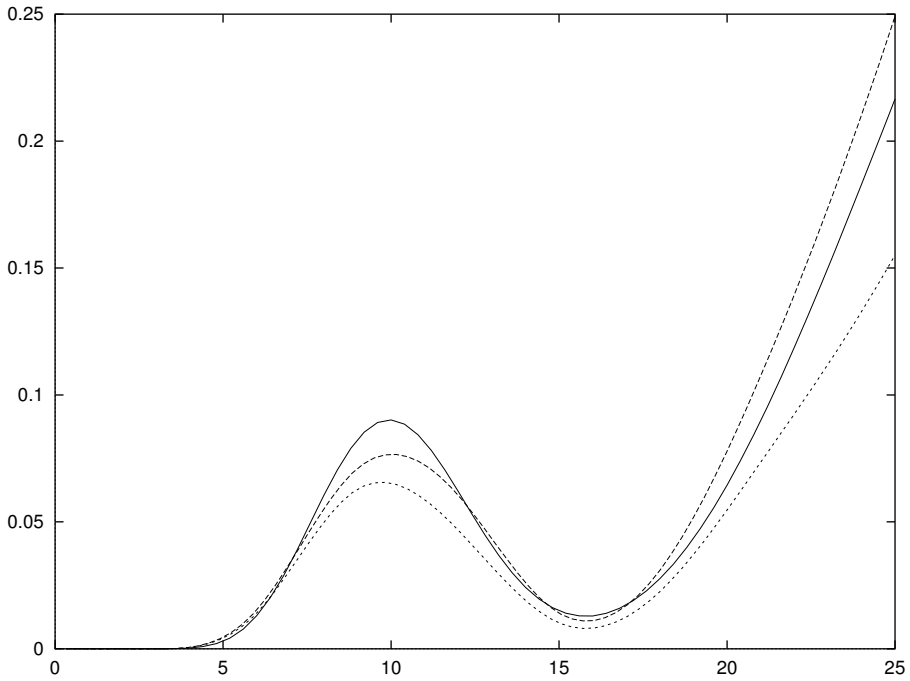

Figure 4

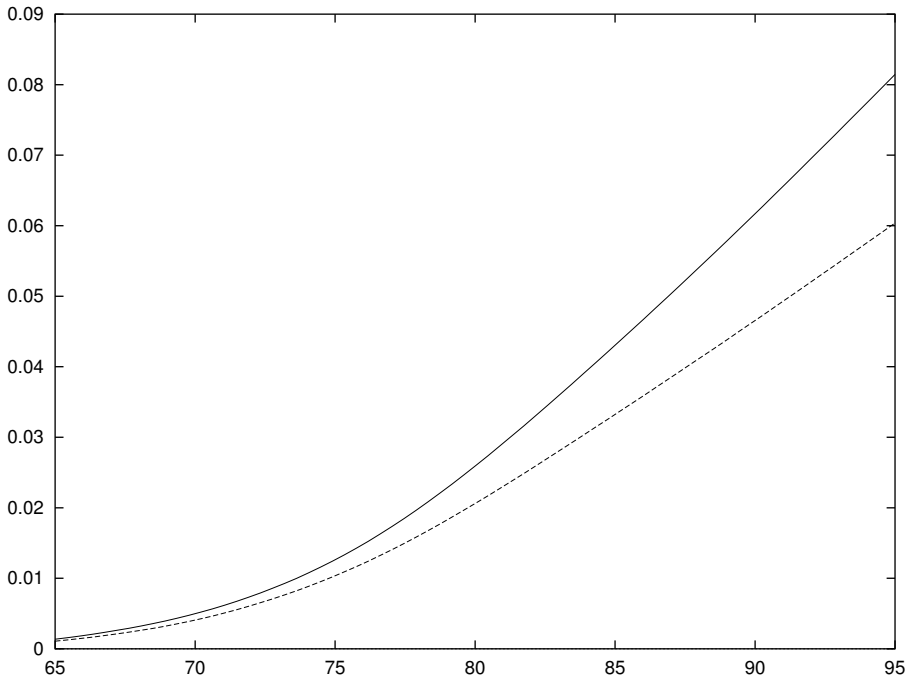

Figure 5

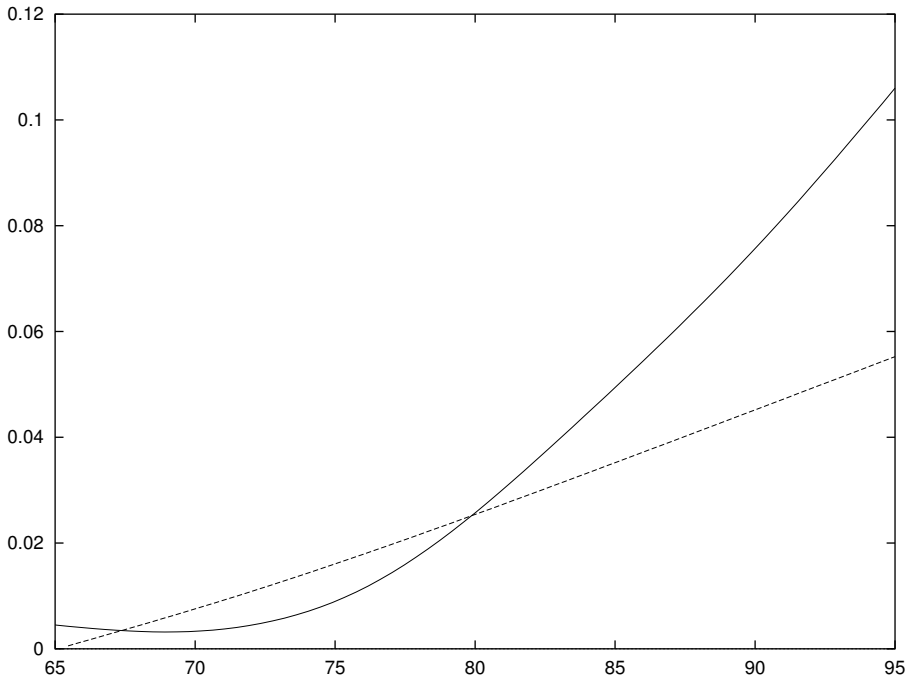

Figure 6

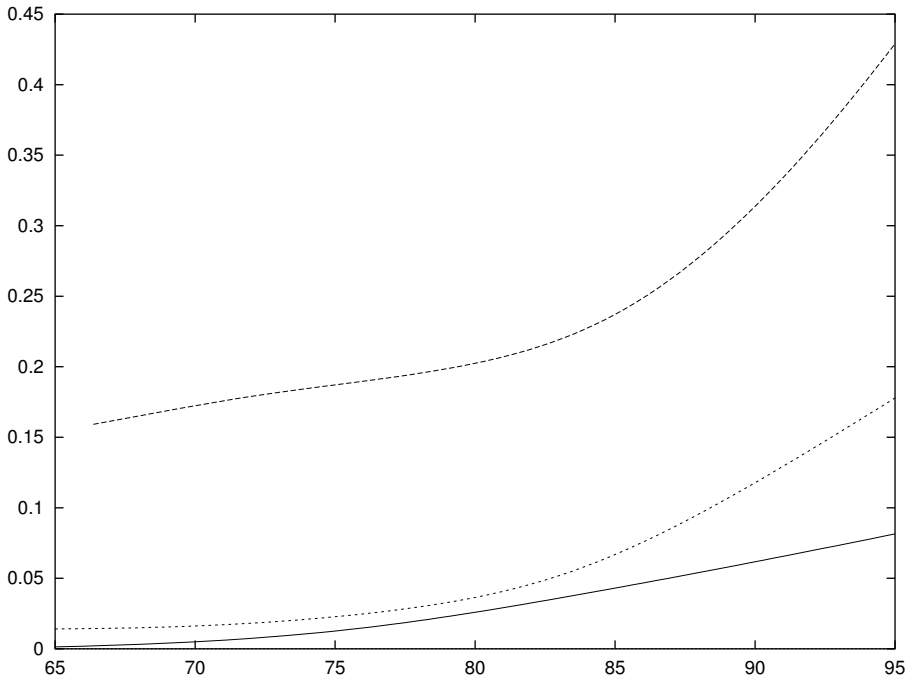

Figure 7