Published in final edited form as:

Biom J. 2010 October ; 52(5): 616–627. doi:10.1002/bimj.200900266.

Multiple Imputation for Estimating the Risk of Developing Dementia and Its Impact on Survival

Binbing Yu^{1,*}, Jane S. Saczynski², and Lenore J. Launer¹

¹Laboratory of Epidemiology, Demography and Biometry, National Institute on Aging, Bethesda, MD 20892

²Division of Geriatrics, the Meyers Primary Care Institute, University of Massachusetts Medical School, Worcester, MA 01605

Summary

Dementia, Alzheimer's disease in particular, is one of the major causes of disability and decreased quality of life among the elderly and a leading obstacle to successful aging. Given the profound impact on public health, much research has focused on the age-specific risk of developing dementia and the impact on survival. Early work has discussed various methods of estimating age-specific incidence of dementia, among which the illness-death model is popular for modeling disease progression. In this article we use multiple imputation to fit multi-state models for survival data with interval censoring and left truncation. This approach allows semi-Markov models in which survival after dementia may depend on onset age. Such models can be used to estimate the cumulative risk of developing dementia in the presence of the competing risk of dementia-free death. Simulations are carried out to examine the performance of the proposed method. We analyze data from the Honolulu Asia Aging Study to estimate the age-specific and cumulative risks of dementia and to examine the effect of major risk factors on dementia onset and death.

Keywords

Competing risk; Dementia; Illness-death model; Interval censoring; Multiple imputation

1 Introduction

Human aging is a worldwide phenomenon, a better understanding of the presentation and impact of dementia is essential to the future of public health (Chapman et al., 2006). As dementia is costly in terms of both personal suffering and economic loss, study of its prevalence, risk factors and potential interventions is emerging as an important facet of public health and health care delivery. The age-specific and cumulative risks of developing dementia and survival after dementia onset are important for decision makers and the general public. These measures can be used to communicate risk estimates, disease burden and comparisons among population subgroups.

The methodology of calculating the risk of developing diseases has been investigated in areas other than dementia. For example, Feuer et al. (1993) and Fay et al. (2003) have discussed the estimation of age-conditional and lifetime risk of developing cancer using cross-sectional cancer incidence and mortality rates from population-based cancer registries.

*Corresponding author: yubi@mail.nih.gov, Phone: +00 301 496 6009, Fax: +00 301 496 4006.

Conflict of interest: None declared.

Because there are still few medications to modulate the course of dementia and the disease may not be recognized by family or physicians in its early stage, there are no reliable registries of dementia cases at present (Launer and Brock, 2004). There are several challenges involved in using data from studies of dementia. First, the subjects are not followed up frequently and the exact ages of dementia onset are not observed, thus creating interval-censored observations. Second, dementia is very rare in younger subjects. Usually the population under study is beyond a certain age, thus creating left-truncated dementia onset times (delayed entry). Third, because of infrequent follow-up exams, subjects may die without their true dementia status being identified. Furthermore, cohorts under study usually consist of very old people who are subject to multiple competing risks of death.

Various methods have been proposed to estimate the age-specific incidence rate of developing dementia in different contexts. Commenges and his group published a series of papers estimating the incidence and mortality rates of dementia. Commenges et al. (1998) and Joly et al. (1998) originally used a penalized likelihood approach to estimate age-specific incidence of dementia. Later, Joly et al. (2002), Commenges et al. (2004), Commenges and Joly (2004) and Commenges and Gegout-Petit (2007) discussed modeling and inferences of Alzheimer's disease and dementia incidence and mortality using an illness-death model under various situations. For the general discussion of the multi-state model (MSM), which includes the illness-death model as a special case, see Andersen and Keiding (2002), Commenges (2002) and Jackson et al. (2003), among others. Their methods were successful and have been widely applied to situations in which dementia cases are identified as a discrete time and death is observed continuously, which is a common design in longitudinal dementia studies. For example, Helmer et al. (2001) examined mortality after dementia and Commenges et al. (2007) discussed the choice between Markov and semi-Markov multi-state models. Commenges and Joly (2004) applied the MSM to three competing events, i.e., dementia, institutionalization and death. Joly et al. (2002) suggested that transition intensity from dementia to death could depend on both current age and dementia onset age, which implied a semi-Markov model.

The parameter estimates of MSMs can be obtained by maximizing the likelihood or penalized likelihood functions (Commenges and Gegout-Petit, 2007). Several statistical packages have been developed to fit MSMs. The R library *tdc.msm* (Meira-Machado et al., 2007) can be used when the disease times are exactly observed or right-censored and the *msm* library (Jackson et al., 2003) can fit a Markov MSM with time-homogenous transition rates. Putter et al. (2007) reviewed statistical methods and practical issues of competing risks data and MSMs. The multiple imputation (MI) method (Rubin, 1987) has been widely used in the analysis for missing data, in particularly for interval-censored survival data (Wei and Tanner, 1991; Pan, 2000). Building on previous work, we apply the MI method to fit the MSM, which is an alternative to the direct maximum-likelihood method. We first impute the disease process, including dementia status and/or dementia-onset age, depending on life history. The imputation creates a complete disease history with exact or right-censored event times. Then standard statistical packages are utilized to obtain the parameter estimates. Covariates can be included in the regression and the models for transition intensities can be either Markov or semi-Markov. Combining the transition intensities from health to dementia and to death, one can estimate the cumulative probabilities of developing dementia in the presence of death.

The rest of the paper is organized as follows. In Section 2, we describe the Honolulu Asia Aging Study and potential bias of conventional analysis. In Section 3, we review the illness-death model and show the likelihood function for survival data with interval-censoring; we then describe the estimation method. In Section 4, we examine the performance of the proposed method by simulation. In Section 5, we analyze data from a population-based

aging study to examine the effect of major risk factors on dementia onset and death. The article ends with discussion in Section 6.

2 Honolulu Asia Aging Study

The Honolulu Heart Program (HHP) is a prospective study of heart disease and stroke for a cohort of Japanese-American men who were born in 1900–1919 and were living in Oahu, Hawaii (Heilbrun et al., 1986). The HHP ended in 1974. As a continuation of the HHP, the Honolulu Asia Aging Study (HAAS) began in 1991 to study dementia prevalence, incidence and risk factors (White, 1992). At the initial wave of the HAAS (called Wave 1), 3,734 people, a subset of surviving HHP subjects, participated in the first examination of cognition function and dementia. Additional examinations were conducted during three follow-up waves: 2 (1994–1996), 3 (1997–1999) and 4 (1999–2000). Dementia screening and diagnosis were conducted by a multistep procedure (White, 1992). Dementia cases diagnosed at the first wave and waves 2–4 were called prevalent and incident cases, respectively. In total, there were 226 prevalent cases at wave 1 and 135, 112 and 52 incident cases at waves 2–4. Death times were recorded exactly through June 30, 2002 regardless of follow-up status.

The follow-up status of the 3508 healthy people in HAAS is shown in Table 1. Among them, 858 subjects did not participate in any follow-up examination for dementia. Using subjects with at least one follow-up exams may introduce bias in two different ways. First, the ages of enrollment at Wave 1 are slightly different for subjects with and without follow-up exams. Since incidence of dementia increases significantly with age, dementia incidence is lower in the younger group. In addition, the subjects without any follow-up die significantly earlier than those with at least one follow-up exam. Furthermore, there is a 2–3 year gap between two consecutive exams. It is possible that some people develop dementia shortly after the previous exam and then die before the next exam. Ignoring the intermittent nature of time of dementia diagnosis may introduce bias (Commenges, 2002).

This analysis concerns two aspects of dementia risk: the risk of developing dementia and the impact of having dementia on survival. Since the baseline mortality rate for younger ages is low, it is hypothesized that dementia would have a more negative impact on survival for younger people. When people get older and have more comorbid diseases, dementia may not be as strong a risk factor for death as are other diseases, e.g., stroke and cancer. We consider the effect of two risk factors: hypertension and education (Launer et al., 1995; Ngandu et al., 2007). The variable for hypertension is medical history of high blood pressure at the beginning of the HHP and the variable for education is total years of education at the first wave of the HAAS. Here, hypertension is defined as a time-independent variable in order to make the interpretation and prediction simple. In the study cohort, the average education is 10.9 years and 15.5% of subjects report hypertension.

3 Models and Estimation

Disease progression can be described by a multi-state process (Commenges, 2002) with a finite number of states $X(t)$ at time $t \geq 0$ and transition probabilities $p_{hj}(s, t, \mathcal{F}(s-)) = P[X(t) = j | X(s) = h; \mathcal{F}(s-)]$, $h = 1, \dots, K$, $j = 1, \dots, K$, where K is the total number of states and $\mathcal{F}(s)$ is the history of the process generated by $\{X(u), u \leq s\}$. When $p_{hj}(s, t, \mathcal{F}(s-))$ is continuous both in s and t , the transition intensity can be defined as

$$\lambda_{hj}(t, \mathcal{F}(t-)) = \lim_{\Delta t \rightarrow 0} \frac{p_{hj}[t, t + \Delta t; \mathcal{F}(t-)]}{\Delta t}. \quad (1)$$

For non-homogeneous Markov processes, we have $\lambda_{hj}(t, \mathcal{F}(t-)) = \lambda_{hj}(t)$. If the transition intensities do not depend on time, i.e., $\lambda_{hj}(t) = \lambda_{hj}$, this is a homogeneous Markov process. For semi-Markov processes, the future process depends not only on the current state, but also on r , the time of entering state h , so the intensity has the general form $\lambda_{hj,r}(t)$.

The MSM is a generalization of the survival process where survival is the ultimate outcome but intermediate states are also of interest (Andersen and Keiding, 2002; Putter et al., 2007). A special class of MSMs, called *illness-death models*, assume that subjects start out as healthy (State 1), and may become ill (State 2) and die (State 3) afterwards, and subjects may die without having the disease of interest (Figure 1). When diseases are recoverable, the transition between State 1 and 2 is bi-directional. In this article, the disease of interest is dementia and the time scale is the age of a subject. We denote dementia onset age by r and age at death by t . Here, we consider a uni-directional model because dementia is an irreversible progressive neurological disorder.

Let Z be the vector of covariates which could be time-dependent. We assume that the hazards for transitions $1 \rightarrow i$ are given by

$$\lambda_{1i}(t|Z) = \lambda_{1i,0}(t) \exp(\beta_{1i}^T Z), i=2, 3, \quad (2)$$

where $\lambda_{12,0}(t)$ and $\lambda_{13,0}(t)$ are baseline hazard functions. The hazard for transition $2 \rightarrow 3$ is

$$\lambda_{23,r}(t|Z) = \lambda_{23,0}(t) \exp(\beta_{23}^T Z + \eta W(r)), \quad (3)$$

where r is the dementia onset age and $W(r)$ is a time-dependent covariate related to r . For instance, if $W(r) = r$, then η represents the effect of dementia onset age on mortality. The term $W(r)$ is not included in the hazards for transitions $1 \rightarrow i$ ($i = 2, 3$). We denote the regression coefficients by $\beta = (\beta_{12}, \beta_{13}, \beta_{23}, \eta)$ and the baseline hazard functions by $\lambda_0 = (\lambda_{12,0}(t), \lambda_{13,0}(t), \lambda_{23,0}(t))$. Various models can be used for the baseline hazards. For example, for the Weibull model, $\lambda_{ij,0}(t) = a_{ij}/b_{ij}(t/b_{ij})^{a_{ij}-1}$, $1 \leq i < j \leq 3$, where a_{ij} is the shape parameter and b_{ij} is the scale parameter.

If the transition hazards $\lambda_{13}(t)$ and $\lambda_{23}(t)$ are proportional, then the hazards λ_{i3} can be written as

$$\lambda_{i3}(t) = \lambda_{3,0}(t) \exp\{\beta_3^T Z + 1(i=2)(\delta + \eta W(r))\}, i=1, 2, \quad (4)$$

where we set $\lambda_{13,0}(t) = \lambda_{23,0}(t) \doteq \lambda_{3,0}(t)$ and $\beta_{13} = \beta_{23} \doteq \beta_3$. Here, δ indicates the additional hazard of being demented on mortality. Joly et al. (2002) showed that the assumption of proportional hazards between $\lambda_{13}(t)$ and $\lambda_{23}(t)$ is reasonable. In general, the assumption of proportional hazards may not hold for all dementia studies and should be validated. The equity of $\beta_{13} = \beta_{23}$ may not be required because they can be estimated separately given observed history of dementia and death.

3.1 Illness-death model for interval-censored and left-truncated data

For simplicity of exposition, we omit Z in the notation of the transition intensities $\lambda_{12}(t|Z)$, $\lambda_{13}(t|Z)$, $\lambda_{23,r}(t|Z)$ and the associated probabilities. For a subject who develops dementia

(State 2) at age r and is still alive at age u , i.e., $R = r$, $T > u$ with $r \leq u$, the conditional probability of being alive with dementia (State 2) at age t (Putter et al., 2007) is

$$P_{22,r}(u, t) = \Pr[T > t | R = r, T > u] = \exp\left(-\int_u^t \lambda_{23,r}(s) ds\right).$$

In a Markov model, $P_{22,r}(u, t)$ does not depend on r because the hazard $\lambda_{23,r}(t)$ is not defined with respect to r . If we have continuous-time observations of the subjects and the dementia onset ages are observed exactly, then the risk of developing dementia could be estimated by a standard MSM.

Let $\Lambda_{1j}(t) = \int_0^t \lambda_{1j}(s) ds$, $j = 2, 3$, and $S_1(t) = \exp\{-\Lambda_{12}(t) - \Lambda_{13}(t)\}$ and let $\mathcal{H}_1(u) = \{R > u, T > u\}$ denote the history of a subject being alive and dementia-free at age u . Following Putter et al. (2007), the likelihood contributions for four possible cases (Joly et al., 2002) are:

- i. The probability of developing dementia during interval $[a, b)$ and still being alive at t ($t \geq b$) is

$$P[a \leq R < b, T > t | \mathcal{H}_1(u)] = \frac{\int_a^b S_1(r) \lambda_{12}(r) P_{22,r}(r, t) dr}{S_1(u)}, \quad (5)$$

- ii. the density of developing dementia in interval $[a, b)$ and dying at age t ($t \geq b$) is

$$P[a \leq R < b, T = t | \mathcal{H}_1(u)] = \frac{\int_a^b S_1(r) \lambda_{12}(r) P_{22,r}(r, t) \lambda_{23,r}(t) dr}{S_1(u)}, \quad (6)$$

- iii. the probability of being alive and dementia free at age t is

$$P[R > t, T > t | \mathcal{H}_1(u)] = \frac{S_1(t)}{S_1(u)}, \quad (7)$$

- iv. and the density of dying at age t and being dementia-free is

$$P[R > t, T = t | \mathcal{H}_1(u)] = \frac{S_1(t) \lambda_{13}(t)}{S_1(u)}. \quad (8)$$

Suppose that the subjects are examined at $\mathcal{T} = \{t_0, t_1, \dots, t_E\}$, where $u = t_0$ is the age of entering in the study and t_E is the age of the last visit. Vital status is observed until age at cut off τ . The age at death (t) is known exactly if $t < \tau$ or is right censored if $t \geq \tau$. If the dementia onset age $r \leq t_E$, then R is interval-censored such that $r \in [a, b]$ with $a, b \in \mathcal{T}$. If $R > t_E$, then the dementia status of the subject is unknown. For subjects that are healthy at age u , the likelihood contributions of the subjects can be calculated using equations (5)–(8) (See details in Section 1 in the Supplementary Document). Given the estimates $\hat{\beta}$ and $\hat{\lambda}_0$, the predicted probabilities in (5)–(8) are obtained by replacing each integral by a sum and by replacing hazard and survival functions by their estimated counterparts. For instance, $\hat{S}_1(t) = \exp\{-\hat{\Lambda}_{12}(t) - \hat{\Lambda}_{13}(t)\}$ and $\hat{\Lambda}_{12}(t) = \sum \hat{\lambda}_{12,0}(t_j)(t_{j+1} - t_j)$, where $t_1, \dots, t_J$ are event ages.

One advantage of using a MSM is that the assumption of independent competing risks is avoided. If in addition we assume that the competing risks of death and dementia are

independent conditional on the covariates, we are able to calculate the cumulative risk of developing dementia since age u as $P[R < \min(b, T) | \mathcal{H}_1(u)]$ (Section 2 in the Supplementary Document).

3.2 Estimation method

In this section, we describe the estimation of MSM using MI. The MI method is a general methodology to handle missing data by data augmentation. Two data augmentation techniques, i.e., the poor man's augmentation (PMDA) and asymptotic normal data augmentation (ANDA), are usually used for interval-censored data (Wei and Tanner, 1991; Pan, 2000). First, we describe the PMDA method. Let superscript l and subscript k denote the l th iteration and k th imputed data sets, respectively. Let $\theta = (\lambda_0, \beta)$ be the parameters to be estimated. The imputation procedure is carried out as follows:

1. Based on the current estimates $\widehat{\theta}^{(l)} = (\widehat{\beta}^{(l)}, \widehat{\lambda}_0^{(l)})$, generate m data sets with dementia status and possibly dementia onset ages. For subjects whose dementia statuses are unknown (Cases 1 and 2 in the supplementary document), first the dementia status is generated from a Bernoulli distribution, and the dementia onset age R is sampled from age range (v, t) , where v is the age at the last exam when the subject was dementia-free and t is the age at death or lost to follow-up. For the subjects who developed dementia in interval (a, b) (Cases 3 and 4 in the supplementary document), the onset age R is sampled from age range (a, b) (see Section 1 of the Supplementary Document).
2. For each complete augmented data set, fit a MSM to obtain the estimates of $\widehat{\beta}_{(k)}^{(l)}$, its covariance $\widehat{\Sigma}_k^{(l)}$ and the baseline hazard functions $\lambda_{0,k}^{(l)}$, for $k = 1, \dots, m$.
3. Update the parameter estimates by letting

$$\widehat{\beta}^{(l+1)} = \frac{1}{m} \sum_{k=1}^m \widehat{\beta}_k^{(l)}, \quad \widehat{\Sigma}^{(l+1)} = \frac{1}{m} \sum_{k=1}^m \widehat{\Sigma}_k^{(l)}, \quad \text{and} \quad \widehat{\lambda}_0^{(l+1)} = \frac{1}{m} \sum_{k=1}^m \lambda_{0,k}^{(l)}.$$

Step (1)–(3) are iterated until the parameters converge. Then the parameters $\widehat{\theta} = (\widehat{\beta}, \widehat{\lambda}_0)$ at convergence are our final estimates. At the final iteration l , the covariance of the final estimates $\widehat{\beta}$ can be approximated by

$$\widehat{\Sigma} = \widehat{\Sigma}^{(l+1)} + \left(1 + \frac{1}{m}\right) \frac{\sum_{k=1}^m [\widehat{\beta}_k^{(l)} - \widehat{\beta}^{(l+1)}]^2}{m-1}.$$

When the amount of missing data is severe, PMDA may underestimate the true variability and the ANDA is preferable (Wei and Tanner, 1991) because the latter is “proper” in the sense of Rubin (1987). For the ANDA method, the posterior distribution of $\widehat{\beta}$ is approximated by a multivariate normal (MVN) distribution,

$$f^{(l+1)}(\beta) = \text{MVN}(\widehat{\beta}^{(l+1)}, \widehat{\Sigma}^{(l+1)})$$

and $\beta_{(k)}^{(l+1)}$ is sampled from $f^{(l+1)}(\beta)$ for each imputation $k = 1, \dots, m$. If a Cox model is used for the hazard functions, the covariance matrix $\hat{\Sigma}$ can be calculated based on partial likelihood and the baseline hazard estimates, given $\hat{\beta}$, can be obtained from the standard Breslow estimator. For a parametric model, both the baseline hazard parameters and the regression coefficients would be drawn from the joint posterior distribution of all parameters. The ANDA imputation method is a Bayesian data augmentation method and produces entire posterior distributions of the parameters, while the PMDA method yields an approximation of the posterior distributions (Wei and Tanner, 1991). Alternatively, one can use the fully Bayesian method for the estimation of unobserved quantities and model parameters.

The initial values $\hat{\beta}_{(0)}$ and $\hat{\lambda}_0^{(0)}$ are estimated using the middle point $(a + b)/2$ for the dementia diagnosis interval. For each augmented dataset, estimates are obtained using standard statistical methods. For example, for fitting a Cox model, one can use the R function `coxph` or the SAS procedure PROC PHREG. Rubin (1987, p114) shows that the efficiency of an estimate based on m imputations is approximately $(1 + \frac{\gamma}{m})^{-1}$, where γ is the rate of missing information. For a rate of missing information as high as 0.9, the relative efficiency of using m imputations, rather than an infinite number of imputations, is 0.92 unit of variance when $m = 10$. So m is set at 10 in the simulation and application.

4 Simulation

We conduct a small simulation to examine the performance of the proposed method. To mimic the real data of a cohort of age 65 or older, we assume that the time scale t is age minus 65. We consider two covariates, where education (Z_1) is related to dementia and hypertension (Z_2) is a risk factor for death. The transition intensity to dementia is

$$\lambda_{12}(t) = \lambda_{12,0}(t) \exp(\beta_1 Z_1),$$

where $\beta_1 = -0.03$ and the baseline hazard $\lambda_{12,0}(t) = a_{12}/b_{12}(t/b_{12})^{a_{12}-1}$ with $a_{12} = 2.15$ and $b_{12} = 29.7$. The corresponding dementia incidence rates are 0.9% when $t = 5$ and 5.9% when $t = 25$ and the subjects with higher education have 3% (relatively) lower incidence rates. The intensities to death are given by

$$\lambda_{i3}(t) = \lambda_{3,0}(t) \exp\{\beta_2 Z_2 + 1(i=2)\delta\}, i=1, 2,$$

where $\beta_2 = 0.2$, $\delta = 2.0$ and the baseline hazard $\lambda_{3,0}(t) = a_3/b_3(t/b_3)^{a_3-1}$ with $a_3 = 2.23$ and $b_3 = 19.5$. Based on these parameters, the median survival times are 17 years for disease-free subjects, 15 years for hypertensive subjects and 7 years for demented subjects, respectively. The distributions of covariates are based on HAAS data. Specifically, education is generated from a normal distribution with a mean of 10.6 and standard deviation of 3.2 and education is truncated at 0. Hypertension is a binary variable with a 20% probability of being positive. The age t_0 of the subjects at the baseline exam follows a Weibull distribution with survival function $S(t_0) = \exp(-(t_0/b)^a)$, where $a = 1.66$ and $b = 7.76$. The resulting median age at baseline is 6.2 years. We assume that the follow-up exams take place every 3 years after the beginning of the study. Thus the dementia onset ages are interval-censored. The study cutoff date is 12 years after the baseline exam. Those subjects with dementia onset or death ages before the baseline exam are excluded from the analysis. To investigate the effect of sample

size, we try sample sizes of 500, 1000 and 4000. For simplicity, we use the ANDA augmentation method. The final results are based on 1000 replications.

Table 2 presents the biases, standard errors of the estimates and actual coverage rates of the 95% confidence intervals (CI) from the simulation. For all three parameters, the empirical biases are relatively small and very similar for the three sample sizes. The estimate of β_1 , representing the effect of education on the risk of dementia onset, has the smallest bias. The standard errors decrease when the sample size increases. The actual coverage rates of the 95% CIs for β_1 and δ are slightly over the nominal level, and those for β_2 are close to the 95% level. Based on this simulation, we see that the proposed method performs well. In practice, the performance of the proposed method depends on sample size, dementia incidence rate, death rate and follow-up exam schedule. For example, if the mortality rate is high compared to the dementia incidence rate, then it is necessary to impute disease history before death.

5 Data Analysis

First, a MSM with Markov assumption is fitted after excluding prevalent cases. For comparison, we also carry out the analysis with prevalent cases included. The covariates include education in years and self-reported hypertension. The reference group consists of subjects with 10.7 years of education and no hypertension. By plotting the estimated hazard functions $\lambda_{13,0}(t)$ and $\lambda_{23,0}(t)$ without the restriction of proportionality, the assumption of proportional hazards is satisfied. We also compared the ANDA and the PMDA imputation schemes and the results were similar (data not shown). Thus, we used the PMDA method for our analysis for its simplicity. However, the ad hoc PMDA imputation method should be used with caution. In all analyses, the calculation stops when $|\hat{\beta}^{(l+1)} - \hat{\beta}^{(l)}| < 0.01$ or the iteration number exceeds 500. The results from different analyses are summarized as follows.

5.1 Markov MSMs with prevalent cases excluded

For the Markov MSMs, both the Weibull model and the Cox model are used for baseline intensities. Table 3 shows the parameter estimates along with standard errors (SE) and Wald p -values from the two models. In the model of transition to death, dementia status is used as a covariate to model the effect of having dementia on the risk of death. The estimates from the Weibull and Cox models are very similar. We see that the coefficient for education in the dementia onset model is -0.034 for the Cox model ($p = 0.020$). This is consistent with research showing that low education might be a risk factor for dementia (Caamaño-Isorna et al., 2006; Ngandu et al., 2007). However, the role of education on dementia risk is controversial and requires more investigation. Hypertension is not significantly associated with dementia incidence, but both hypertension and dementia significantly increase the death rate.

Figure 2 shows age-specific hazard rates and cumulative probabilities of dementia and death for the reference group. The smoothed curves are estimated from the Weibull model, and the dots are from the Cox model. In Figure 2a, the age-specific dementia incidence rate from the Weibull model is 6.1% for age 95, which is close to the estimate of 5.5% for age 95 based on the Paquid research program (Joly et al., 2002). Figure 3 shows the crude cumulative probabilities of developing dementia by hypertension status with death as a competing risk for a 70-year subject in the reference group. For subjects with and without hypertension, the cumulative risks of developing dementia are 22.1% and 25.9%, respectively. This contra-intuitive finding shows the effect of the competing risk of death on the probability of developing dementia. Although hypertension increases dementia incidence, it much more

greatly impacts the risk of death, so that subjects with hypertension tend to die before they can develop dementia.

5.2 Semi-Markov MSMs

To examine whether early onset of dementia imposes a higher risk of death, we fit a semi-Markov MSM. Because the Cox and Weibull models for baseline intensities yield similar results, we only present the results from the Cox model. To examine the effect of including prevalent dementia cases, we carry out the analysis with prevalent cases included and excluded. The parameter estimates and the SEs are shown in Table 4.

When the prevalent cases are excluded, the regression coefficients for education and hypertension are very close to those from the Markov model in Table 3. Having dementia increases the hazard of death 7.9 ($= e^{2.070}$) fold ($p < 0.001$). The coefficient for dementia onset age is -0.115 ($p < 0.001$). This means that for demented subjects, onset one year later will decrease the hazard of death by $\exp(-0.115) - 1 = 10.9\%$. This is reasonable because dementia is a slowly progressing disease and older people tend to die from other diseases. For people in very old age groups, dementia coexists with many comorbid conditions, e.g., stroke, heart failure and cancer, and may not be as serious a risk factor for death compared to other disease conditions. When the prevalent cases are included, we assume that these subjects were dementia-free at the end of the HHP. The prevalent cases are handled as interval censoring between the end of the HHP and the first wave of the HAAS. Overall, the parameter estimates from using and not using the prevalent cases are very close, except that the coefficients of hypertension increase slightly from 0.205 to 0.253 for dementia and decrease from 0.212 to 0.123 for death. This may indicate a stronger association between hypertension and early-onset dementia. In addition, SEs for dementia onset parameters are smaller when prevalent cases are included. This is because of the larger sample size when prevalent cases are included.

For comparison, we show the cumulative hazard rates of dementia and death for the reference group in Figure 4. The cumulative hazard rates are defined as $\Lambda_{1j}(t) = \int_0^t \lambda_{1j}(s) ds$. In Figure 4a, the cumulative hazard rate of dementia with prevalent cases included is slightly higher than that with prevalent cases excluded before age 75, but is lower after age 75. This is because subjects with incident dementia are older and contribute more to late-onset dementia. The dementia incidence rate is so low before age 65 that the absolute difference between the two approaches is very small before age 75. In Figure 4b, we see that the cumulative hazard rates of death with prevalent cases included and excluded are very close.

By including the prevalent cases, one can estimate dementia incidence rates before age 70. In contrast, when the prevalent cases are excluded, the dementia incidence rate before age 70 is not estimable because all subjects are older than 71. However, there are some caveats to using prevalent cases. Since the HHP ended in 1974 and the HAAS began in 1991, there is a long gap time of 28 years. The prevalent cases are much younger at their dementia onset and have to survive long enough to be included in the HAAS. Information on death and dementia status cannot be tracked for the individuals who died before the HAAS. Among them, some may have developed dementia and died. Thus, the prevalent cases have a lower death rate and may be a biased sample of all dementia cases. In addition, the imputation of dementia status assumes a constant hazard during the dementia onset intervals. Analysts should take caution when using prevalent cases with such a long gap time. Likelihood expressions were already described by Keiding (1991) and used by Saarela et al. (2009) in a similar context. The approach therein is conditional on survival until the study onset. The probability-based approach to the illness-death model in this paper is not new, but handles dementia incidences as interval-censoring.

Members of the original HHP cohort who died before the start of HAAS were not included in the present analysis, so dementia cases are over-represented with respect to the population at risk. This may partly explain the observed diluted association of hypertension on death and higher cumulative incidence of dementia with prevalent dementia cases included. However, the HAAS cohort was only a subset of the survivors of the HHP cohort and we had no information on disease history and vital status for individuals who were not in HAAS. Thus, including the prevalent cases in the analysis will not be helpful since they would have no likelihood contribution. This alternative approach would be to condition on survival until the end of HAAS, however, it does not fit directly into the interval censoring framework described in the present paper.

6 Discussion

As a simple alternative to previously proposed penalized likelihood estimation methods (Joly et al., 2002), the multiple imputation approach allows the use of standard Cox regression techniques. Importantly, valid variance estimators can be easily obtained following the formulas of Rubin (1987).

We only used hypertension at a fixed time point. However, hypertension is a dynamic process and the effect of hypertension on dementia and mortality is complex. Strictly speaking it should be treated as a time-dependent variable. Estimation and prediction based only on a fixed history of hypertension is limited in scope in this article.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The research was supported in part by the Intramural Research Program of the National Institute on Aging. The simulation used the high-performance computational capabilities of the Biowulf Linux cluster at the National Institutes of Health (<http://biowulf.nih.gov>). The authors wish to thank two anonymous reviewers for their helpful comments.

References

- Andersen PK, Keiding N. Multi-State Models for Event History Analysis. *Statistical Methods in Medical Research*. 2002; 11:91–115. [PubMed: 12040698]
- Caamaño-Isorna F, Corral M, Montes-Martínez A, Takkouche B. Education and dementia: a meta-analytic study. *Neuroepidemiology*. 2006; 26:226–232. [PubMed: 16707907]
- Chapman DP, William SM, Strine TW, Anda RF, Moore MJ. Dementia and its implications for public health. *Preventive Chronic Disease*. 2006; 3:1–13.
- Commenges D, Letenneur L, Joly P, Alioum A, Dartigues JF. Modeling age-specific risk: application to dementia. *Statistics in Medicine*. 1998; 17:1973–1988. [PubMed: 9777690]
- Commenges D, Joly P, Letenneur L, Dartigues JF. Incidence and mortality of Alzheimer's disease or dementia using an illness-death model. *Statistics in Medicine*. 2004; 23:199–210. [PubMed: 14716722]
- Commenges D, Joly P. Multi-state model for dementia, institutionalization, and death. *Communication in Statistics, Theory and Methods*. 2004; 33:1315–1326.
- Commenges D, Gegout-Petit A. Likelihood for generally coarsened observations from multistate or counting process models. *Scandinavian Journal of Statistics*. 2007; 34:432–450.
- Commenges D. Inference for multi-state models from interval censored data. *Statistical Methods in Medical Research*. 2002; 11:167–182. [PubMed: 12040695]

- Commenges D, Joly P, Gegout-Petit A, Liqueur B. Choice between semi-parametric estimators of Markov and non-Markov multi-state models from coarsened observations. *Scandinavian Journal of Statistics*. 2007; 34:33–52.
- Fay MP, Pfeiffer R, Cronin KA, Le C, Feuer EJ. Age conditional probabilities of developing cancer. *Statistics in Medicine*. 2003; 22:1837–1848. [PubMed: 12754719]
- Feuer EJ, Wun LM, Boring CC, Flanders WD, Timmel MJ, Tong T. The lifetime risk of developing breast cancer. *Journal of the National Cancer Institute*. 1993; 85:892–897. [PubMed: 8492317]
- Heilbrun LK, Kagan A, Nomura A, Wasnich RD. The origins of epidemiologic heart diseases, cancer and Osteoporosis among Hawaii Japanese. *Hawaii Medical Journal*. 1986; 44:294–296. [PubMed: 4055356]
- Helmer C, Joly P, Letenneur L, Commenges D, Dartigues JF. Mortality with dementia: results from a French prospective community-based cohort. *American Journal of Epidemiology*. 2001; 154:642–648. [PubMed: 11581098]
- Jackson CH, Sharples LD, Thompson SG, Duffy SW, Couto E. Multi-state Markov models for disease progression with classification error. *The Statistician*. 2003; 52:193209.
- Joly P, Commenges D, Letenneur L. A penalized likelihood approach for arbitrarily censored and truncated data: application to age-specific incidence of dementia. *Biometrics*. 1998; 54:185–194. [PubMed: 9574965]
- Joly P, Commenges D, Helmer C, Letenneur L. A penalized likelihood approach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*. 2002; 3:433–443. [PubMed: 12933608]
- Keiding N. Age-specific incidence and prevalence: a statistical perspective. *Journal of the Royal Statistical Society, Series A*. 1991; 154:371–412.
- Launer LJ, Masaki K, Petrovitch D, Foley D, Havlik RJ. The association between midlife blood pressure levels and late-life cognitive function. The Honolulu-Asia Aging Study. *Journal of the American Medical Association*. 1995; 274:1846–1851. [PubMed: 7500533]
- Launer LJ, Brock DB. Population-based studies of AD: message and methods: an epidemiologic view. *Statistics in Medicine*. 2004; 23:191–197. [PubMed: 14716721]
- Meira-Machado L, Cadarso-Suárez C, De Uña-Álvarez J. tdc.msm: An R library for the analysis of multi-state survival data. *Computer Methods and Programs in Biomedicine*. 2007; 86:131–140. [PubMed: 17350136]
- Ngandu T, von Strauss E, Helkala EL, Winblad B, Nissinen A, Tuomilehto J, Soininen H, Kivipelto M, et al. Education and dementia: what lies behind the association? *Neurology*. 2007; 69:1442–1450. [PubMed: 17909157]
- Pan W. A multiple imputation approach to Cox regression with interval censored data. *Biometrics*. 2000; 56:199–203. [PubMed: 10783796]
- Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*. 2007; 26:2389–2430. [PubMed: 17031868]
- Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. Wiley; New York: 1987.
- Saarela O, Kulathinal S, Karvanen J. Joint analysis of prevalence and incidence data using conditional likelihood. *Biostatistics*. 2009; 10:575–587. [PubMed: 19458343]
- Teng EL, Hasegawa K, Homma A, et al. The Cognitive Abilities Screening Instrument (CASI): a practical test for cross-cultural epidemiological studies of dementia. *International Psychogeriatrics*. 1992; 6:45–58. [PubMed: 8054493]
- Wei GCG, Tanner MA. Application of multiple imputation to the analysis of censored regression data. *Biometrics*. 1991; 47:1297–1309. [PubMed: 1786320]
- White LR. Toward a program of cross-cultural research on the epidemiology of Alzheimer's disease. *Current Science*. 1992; 63:456–469.

Fig. 1.

The illness-death model: λ_{12} , λ_{13} , and $\lambda_{23,r}$ are the intensity functions

a. Dementia incidence rate

b. Mortality rate

Fig. 2.
Age-specific incidence rates of dementia and death

Fig. 3.
Cumulative probability of developing dementia

a. Cumulative hazard rate of dementia $\Lambda_{12}(t)$ b. Cumulative hazard rate of death $\Lambda_{13}(t)$

Fig. 4.
Cumulative hazard rates of dementia incidence and death

Table 1
Dementia diagnosis and follow-up status for participants free of prevalent dementia in HAAS

Participation status				Dementia diagnosis			
Wave 2	Wave 3	Wave 4	All	Not diagnosed	Wave 2	Wave 3	Wave 4
No	No	No	858	858	.	.	.
		Yes	22	21	.	.	1
	Yes	No	16	12	.	4	.
		Yes	20	16	.	3	1
Yes	No	No	648	579	69	.	.
		Yes	24	22	.	.	2
	Yes	No	477	406	34	37	.
		Yes	1443	1295	32	68	48

Table 2

Summary of the parameter estimates from simulation

Parameter (true value)	Sample size	Bias	Standard error	Coverage rate
$\beta_1(-0.03)$	500	0.001	0.057	97.0%
	1000	-0.007	0.040	96.8%
	4000	-0.000	0.020	97.7%
$\beta_2(0.2)$	500	-0.013	0.157	95.5%
	1000	-0.010	0.111	94.4%
	4000	-0.010	0.055	95.5%
$\delta(2)$	500	-0.103	0.198	96.9%
	1000	-0.094	0.136	95.7%
	4000	-0.091	0.089	95.5%

Table 3

Parameter estimates from the Markov MSMs

Transition	Parameter	Weibull		Cox	
		Estimate (SE)	p-value	Estimate (SE)	p-value
Onset (1→2)	Education	-0.033 (0.017)	0.057	-0.034 (0.015)	0.020
	Hypertension	0.064 (0.155)	0.677	0.158 (0.150)	0.292
	log(scale)	3.385 (0.035)	< .001		
	log(shape)	0.795 (0.055)	< .001		
Death (1, 2 →3)	Education	-0.007 (0.008)	0.400	-0.004 (0.008)	0.582
	Hypertension	0.215 (0.064)	0.001	0.208 (0.065)	0.001
	Dementia(δ)	0.799 (0.081)	< .001	0.843 (0.078)	< .001
	log(scale)	3.038 (0.015)	< .001		
	log(shape)	0.806 (0.030)	< .001		

Table 4

Parameter estimates from the Cox semi-Markov MSM

Transition	Parameter	Exclude prevalent cases		Use prevalent cases	
		Estimate (SE)	p-value	Estimate (SE)	p-value
Onset (1 → 2)	Education	-0.034 (0.015)	0.023	-0.030 (0.012)	0.010
	Hypertension	0.205 (0.143)	0.151	0.253 (0.103)	0.014
Death (1 → 3)	Education	-0.003 (0.008)	0.665	-0.000 (0.007)	0.960
	Hypertension	0.212 (0.065)	0.001	0.123 (0.061)	0.044
	Dementia status (δ)	2.070 (0.205)	< .001	2.219 (0.341)	< .001
	Onset age (η)	-0.115 (0.017)	< .001	-.053 (0.011)	< .001