Published in final edited form as:

Comput Stat Data Anal. 2010 August 1; 54(8): 1921–1929. doi:10.1016/j.csda.2010.02.025.

A Bayesian MCMC approach to survival analysis with doubly-censored data

Binbing Yu

Laboratory of Epidemiology, Demography and Biometry, National Institute on Aging, National Institutes of Health, Bethesda, MD 20892, U.S.A.

Abstract

Doubly-censored data refers to time to event data for which both the originating and failure times are censored. In studies involving AIDS incubation time or survival after dementia onset, for example, data are frequently doubly-censored because the date of the originating event is interval-censored and the date of the failure event usually is right-censored. The primary interest is in the distribution of elapsed times between the originating and failure events and its relationship to exposures and risk factors. The estimating equation approach [Sun, et al. 1999. Regression analysis of doubly censored failure time data with applications to AIDS studies. *Biometrics* 55, 909-914] and its extensions assume the same distribution of originating event times for all subjects. This paper demonstrates the importance of utilizing additional covariates to impute originating event times, i.e., more accurate estimation of originating event times may lead to less biased parameter estimates for elapsed time. The Bayesian MCMC method is shown to be a suitable approach for analyzing doubly-censored data and allows a rich class of survival models. The performance of the proposed estimation method is compared to that of other conventional methods through simulations. Two examples, an AIDS cohort study and a population-based dementia study, are used for illustration. Sample code is shown in the appendix.

Keywords

AIDS; dementia; doubly censored data; incubation period; MCMC; midpoint imputation

1. Introduction

Doubly-censored data often arise in epidemiologic studies of disease progression in which the originating event time U and the subsequent failure event time V could be exact, right-censored or interval-censored. In these studies, the object of interest is the survival function $S(t)$ of the elapsed time between the two events, that is, $T = V - U$.

Sampling schemes involving doubly-censored data can occur when the originating event is not directly observable but is detected via periodic screening. The originating event time U is known only within an interval $[L, R]$. For example, in AIDS epidemiologic cohort studies, one common research topic is to estimate the AIDS incubation time, the gap time between human immunodeficiency virus (HIV) infection (U) and the diagnosis of AIDS (T). HIV infection time is interval-censored due to recruitment of HIV positive patients and/or determination of

infection times through periodic screening of blood samples (De Gruttola and Lagakos, 1989). In addition, the AIDS diagnosis times are often right-censored due to incomplete followup. Another typical situation is the cohort survival study of dementia patients, in which the goal is to estimate survival time from dementia onset (U) to death (V). As the study population is recruited and periodically examined for dementia, the dementia onset time is interval-censored: the onset time is only known to have happened between two consecutive examinations. The death time for demented may be right censored because of dropout or deaths from causes other than dementia. Thus, survival time after dementia onset is also doubly-censored.

A popular ad-hoc method is to impute the originating event time by the midpoint of the interval. Law and Brookmeyer (1992) showed that the midpoint method yields biased estimates of survival function $S(t)$ and underestimates the variances. Since Turnbull (1976) proposed a self-consistency non-parametric estimate of survival function for arbitrarily censored or truncated survival data, there has been considerable research effort in this area. Sun (1999) proposed an estimating equation approach for regression analysis. Later, Pan (2001) extended the method using multiple imputation. The multiple imputation approach, which is referred to as the SP method, can be briefly described as follows: First, $\hat{H}$, the estimate of distribution function $\hat{H}$ for U , is computed. Then multiply-imputed originating event times, $U_{(1)}, \dots, U_{(M)}$, are generated independently from $\hat{H}$, conditional on $L \leq U_{(m)} \leq R$. Finally the parameter estimates are calculated by summarizing estimates for each imputed data set.

There are several potential issues about the SP method. First, it assumes that the function H is known and can be replaced by its estimate $\hat{H}$. Therefore the variation in the estimate $\hat{H}$ is ignored. As a remedy, Pan (2001) suggested using a Bayesian bootstrap. Second, the distribution of U is assumed to be the same for all subjects regardless of individual characteristics. In reality, the HIV infection time may be related to individual sexual behavior, drug use and socioeconomic status; dementia onset age can be influenced by education level and mid-life blood pressure (Launer et al. 1998). Third, survival time T is assumed to be independent of originating event time U . This may not be true even after adjusting for covariates. For example, early onset dementia cases are usually more severe and have higher mortality. Survival after dementia onset is expected to be related to dementia onset age (Wolfson et al. 2001; Xie et al. 2008).

Kim et al. (1993) and Goggins et al. (1999) proposed a full likelihood method by modeling the joint distribution of both originating and failure event times. The full-likelihood method is likely to be more efficient asymptotically if the model is specified correctly. However, it is also criticized because of the great computation effort involved as well as non-identifiability problems due to a large number of parameters. The Bayesian Markov chain Monte Carlo (MCMC) method is an alternative to the full likelihood method and is especially appropriate for doubly-censored data. Kneib (2006) proposed additive and geoadditive regression models for interval-censored data and Zhang et al. (2008) used a hierarchical Cox model for doubly censored data from multiple studies.

Building on previous work, we discuss the use of the Bayesian MCMC method for doubly-censored data. First we demonstrate the importance of incorporating covariates into the distribution of originating events so as to increase accuracy in prediction. Second, we show that the Bayesian approach allows flexible survival models for the elapsed time; these models could be proportional or additive hazards models or cure models. The rest of this article is organized as follows. The statistical model and computation methods are described in Section 2. Simulations are conducted to compare the performance of available methods versus the proposed method in Section 3. Then the proposed method is applied to data from two cohort studies in Section 4. The final discussion is given in Section 5.

2. A Bayesian MCMC Approach for Doubly-Censored Data

2.1. Data structure

Suppose that there are n subjects in the study. For the i th subject, the originating event time is denoted by U_i and the latent failure time is denoted by V_i^* . The time V_i^* may be right censored with C_i so that we observe $V_i = \min(V_i^*, C_i)$ and $\delta_i = I(V_i^* \leq C_i)$ instead. Due to interval censoring, we do not observe U_i directly, but only know that U_i lies in an interval $[L_i, R_i]$. In addition, U_i may be left-truncated by the delayed entry time A_i if the originating events already occurred for some individuals. The analysis goal is to investigate the relationship of elapsed time, $T_i = V_i^* - U_i$, with a vector of covariates X_i . In the analysis, we focus on subjects who already experience the originating event, but have not had the failure event. The subjects who have experienced both events are excluded from the sample, which may lead to loss of information. In the usual situations, the proportion of subjects with both events is relatively small.

2.2. Models and likelihood

Two extensions of the current methods are possible. First, flexible survival models, including parametric cure models, additive and proportional hazards models, can be used for the distribution of elapsed time. Let $\lambda(t|X)$ and $S(t|X)$ be the hazard and survival functions of elapsed time T given covariates X . For example, the proportional hazards model specifies the hazard function as

$$\lambda(t|X) = \lambda_0(t) \exp(\beta'X), \quad (1)$$

where β is the vector of regression coefficients and the baseline hazard function $\lambda_0(t)$ can be unspecified in the Cox semiparametric model or can be piecewise constant with $\lambda_0(t) = \lambda_j \mathbf{I}(t_{j-1} < t \leq t_j)$ for $j = 1, \dots, J$. If cure exists for the disease or there are long-term survivors, a cure model is often used. Initially proposed by Yakovlev et al. (1993) and extended by Chen et al. (1999), a commonly used cure model assumes a survival function of the form

$$S(t|X) = \exp(-\pi F(t)), \quad (2)$$

where $\pi = \exp(\beta'X)$ is the parameter related to cure rate and $F(t)$ is a distribution function.

Second, the parameter estimation for $S(t)$ can be improved by relating originating event times to covariates. Let $H(u|X)$ and $h(u|X)$ be the survival and density functions for U , respectively. For example, the lognormal (LN) model assumes that

$$H(u|X) = 1 - \Phi\left(\frac{\log u - \alpha'X}{\sigma}\right), \quad (3)$$

where α is the vector of regression coefficients.

If the originating event is observed exactly as $U = u$, the likelihood contribution is

$$l(\theta|u, A_i, X_i, V_i, D_i, \delta_i) = \frac{h(u|X_i)}{H(A_i|X_i)} \lambda(V_i - u|X_i)^{\delta_i} S(V_i - u|X_i), \quad (4)$$

where θ denotes the vector of all parameters. In this situation, H and (β, λ_0) can be estimated separately by factoring the likelihood function (4). Because U is interval-censored, the likelihood function for observed data takes an integral form

$$L(\theta) \propto \prod_{i=1}^n \left(\int_{L_i}^{R_i} l(\theta|u, A_i, X_i, V_i, D_i, \delta_i) du \right). \quad (5)$$

The originating event time U , the elapsed time T and the covariates are modeling jointly in the likelihood functions (4) and (5). The likelihood functions do not require independence between U and T . In the functions $\lambda(t|X)$ and $S(t|X)$, the originating event time can be used as a covariate. Generally, there is no closed form for the likelihood function $L(\theta)$. A direct maximization method, e.g., Newton-Raphson method or the Expectation-Maximization (EM) algorithm, can be used to find the maximum likelihood estimates (MLEs).

2.3. Bayesian MCMC analysis

The Bayesian MCMC method is a useful alternative to obtain the posterior distributions of the parameters. The asymptotic properties of the MLEs can be derived from the Central Limit Theorem and the inference of the parameters for the Bayesian method can be based on the posterior distributions.

The Bayesian MCMC method for the doubly-censored data is described as follows: First, the distributions for the originating event and elapsed times are specified as described in Section 2.2. Second the priors of the unknown parameters are assigned. We assume the elements of θ are independent of each other. The priors of regression coefficients α and β are multivariate normal (MVN), i.e., $\alpha \sim \text{MVN}(0, \Sigma_\alpha)$, $\beta \sim \text{MVN}(0, \Sigma_\beta)$, the priors for variance-covariance matrix is $\Sigma_\beta \sim \text{Wishart}(v_\beta, \Gamma)$, the prior for variance parameter is $\sigma^2 \sim \text{Inverse Gamma}(a, b)$ and $r \sim \text{Exponential}(c)$. Hyperparameters are chosen such that the priors of the parameters tend to be weakly informative. For example, we set $a = b = c = 0.001$. This choice implies that the a prior mean equals 1 and variance equals 1000 for the parameter σ^2 . Parameter estimates and the 95% credible intervals (CI) can be obtained from the MCMC samples. Some caution should be exercised because the choice of hyperparameters may affect the final parameter estimates.

The implementation of this method is relatively easy in the publicly available software WinBUGS (Spiegelhalter et al. 2005). Initial values for the parameters are selected by starting with the prior means and covering ± 2 standard deviations. In order to see how stable the final estimates are, multiple parallel chains should be carried out. In the simulation and applications, we ran two parallel chains with different initial values. For each chain, the 2,000 burn-in samples were excluded and the results were based on the additional 5,000 samples. The convergence of the MCMC samples of parameters θ can be assessed using the criteria of Geweke (1992), Heidelberger and Welch (1983) and Gelman and Rubin (1992). Sample WinBUGS code is shown in the appendix.

3. Simulation Study

A small simulation study was conducted to compare the performance of the exact, midpoint, SP methods and the proposed Bayesian method. A binary covariate x was used to indicate group membership and each group consisted of 100 subjects. The originating event time was

generated from two scenarios: a common distribution $LN(3.7, 0.3^2)$ for all subjects or a distribution $LN(3.8 - 0.4x, 0.3^2)$ dependent on covariate x . We assumed that every subject was followed up annually and the originating event time was interval-censored between two consecutive visits. Whether or not a subject completed each annual follow-up was generated from an independent Bernoulli variable with probability $p = 0.5$ or 0.3 to create moderate or heavy censoring. The average widths of the censoring intervals were 2.9 and 5.5 for moderate and heavy censoring, respectively. The elapsed time was simulated from a Weibull distribution with shape parameter $a = 2.3$ and scale parameter $b = \exp(\beta_0 + \beta_1 x)$, where $\beta_0 = 4.3$ and $\beta_1 = 0.3$. Resulting failure times were subject to about 10% random right censoring. Here we focused on parameter β_1 , which represents the effect of x on elapsed time. We compared the bias, standard error (SE) and coverage rate (CR) of the 95% confidence/credible intervals (CI) from four different methods. For the exact, midpoint and the SP methods, a semiparametric Cox model was used so the baseline hazard was unspecified. For the Bayesian method, a piecewise constant baseline hazard with annual intervals was assumed.

Results are presented in Table 1. When the distribution of originating event times does not depend on the covariate, all four methods, except for the midpoint method, are similar. The midpoint method yields larger biases and it underestimates the variance of the parameters, thus leading to a CR lower than the 95% level. The actual CRs for the other three methods are all close to the nominal 95% level. The SEs from the SP and Bayesian methods are larger than those from the other two methods. This makes sense because originating event times are assumed to be observed exactly and fixed for the exact and midpoint methods. The SEs for the heavy censoring are slightly larger than those for the moderate censoring.

When originating event time is covariate-dependent, performance of the SP method gets worse while performance of the other three methods remains similar. The exact method still has minimal bias and the best coverage rate. The biases of the parameters from the midpoint method are larger than those from the other three methods. Especially when there is heavy censoring, the Bayesian method is better than the midpoint and the SP methods. In summary, when the distribution of originating event time depends on the covariate, the SP method does not incorporate the additional information and may lead to biased estimates. This illustrates the importance of utilizing available information to impute originating event time, as does the proposed Bayesian method.

4. Applications

4.1. Honolulu Asia Aging Study

Survival after dementia onset has been investigated in cohort studies in various settings. The median survival time ranges from 3 to 9 years. Variation in survival might be caused by differences in study designs. Some studies are population-based, some are community-based, and others are institution-based. Most studies are based on prevalent cases (Wolfson et al. 2001), which could lead to bias because of the exclusion of patients with relatively rapid progressive dementia. Analysis using incident dementia avoids this bias (Xie et al. 2008), but introduces other difficulties such as how to decide the time of onset. To date, most analyses of dementia survival use different strategies to impute dementia onset times. These strategies are based on dates of first symptoms and first consultation, on the middle point between two examinations, or on initial clinical diagnosis (Larson et al. 2004). Waring et al. (2005) used different definitions of onset times and reported a median survival of 5.7 years from diagnosis of dementia, but 10.5 years from onset of the first reported symptomatic sign.

Started in 1991, the Honolulu-Asia Aging Study (HAAS) was a population-based aging study to examine dementia prevalence, incidence and risk factors for a cohort of 3734 Japanese-American men in the island of Oahu, Hawaii (Heilbrun et al. 1986). In the beginning of the

study, all the participants were least 71 years old. Assessment of cognitive function and dementia was assessed during four examinations in 1991, 1994, 1997 and 2000. The average gap time between two consecutive exams was 3 years. Dementia cases in HAAS were classified into pure Alzheimer's disease (AD), pure vascular dementia (VD), AD with VD, and other dementia (White et al. 1996). At baseline examination 226 prevalent cases of dementia were diagnosed and in the three followup examinations, 135, 112 and 52 incident cases were diagnosed. We considered three covariates: years of education (edu), Apolipoprotein E $\epsilon 4$ genotype ($ApoE$) and diabetes status. For the cohort under study, the average education was 10.5 years and 18% were ApoE $\epsilon 4$ carriers. Among the 299 incident dementia cases, 63% died before the end of the study.

We ran 2 parallel chains with 2,000 burn-ins and the parameter estimates were based on the additional 5000 MCMC runs. Convergence was achieved for all parameters based on the rules by Gelman and Rubin (1992) and Heidelberger and Welch (1983). The parameter estimates are not sensitive to the choice of hyperparameters and initial values. A lognormal model (3) was used for dementia onset age with left truncation, so a negative coefficient α meant earlier dementia onset. To examine the goodness of fit of the lognormal model empirically, we compared the survival curves of dementia onset age from the lognormal model and the Kaplan-Meier method using the midpoint in Figure 1a. The Kaplan-Meier curve is based on all subjects without adjusting for covariates and the lognormal survival curve is calculated at the centered values of the covariates. We see that the survival functions from both methods were very close to each other. A piecewise constant hazard was used for the baseline hazard from survival time after dementia. Parameter estimates of the dementia onset age and survival after dementia are shown in Table 2. We see that patients with diabetes or lower education have a shorter survival time. Subjects with ApoE $\epsilon 4$ genotype have earlier dementia onset. However, ApoE $\epsilon 4$ carriers live longer than non-carriers after dementia onset. Similar finding has been reported that Alzheimer's disease patients carrying ApoE $\epsilon 4$ have a lower rate of deterioration and higher survival rate than patients without ApoE $\epsilon 4$ [15]. Patients with pure VD have a significantly shorter survival time than those with the other three types of dementia. Based on these estimates, the median survival time after dementia onset is 5.6 years with a 95% CI 4.5–6.1 years for AD. The median survival times are 5.8 (4.6–6.3) years for AD/VD mixed dementia, 4.8 (3.5–5.1) years for pure VD and 5.4 (4.4–7.9) years for other dementia, respectively.

As a comparison, Figure 2 shows the baseline cumulative hazards from the Cox model with midpoint imputation and the Bayesian model with piece-wise constant baseline hazards. We see that the hazard functions from both methods are very close. The parameter estimates and SEs from the Cox model with midpoint imputation are $\beta_1 = 0.020$ (0.022) for education, $\beta_2 = -0.238$ (0.192) for ApoE $\epsilon 4$ and $\beta_3 = 0.337$ (0.202) for diabetes. Parameter estimates from the two models are slightly different, mainly because the Bayesian model assumes that dementia onset time is related to the covariates while the midpoint method ignores this information.

4.2. AIDS Incubation Time

Many authors have considered a cohort study of hemophiliacs who were treated in two hospitals in France (Kim et al., 1993; Sun et al., 1999; Goggins et al., 1999). These patients were at risk for HIV infection due to contaminated blood. At the end of study 188 patients were HIV-infected, and the infection times were interval-censored. Among them, 41 progressed to AIDS or related symptoms.

A Kaplan-Meier analysis of AIDS incubation time showed that more than 50% of the patients with HIV infections did not develop AIDS in the end of the study. Thus, we used a Bayesian cure model (Chen et al. 1999) to model the long-term survivors of developing AIDS. Patients had been classified into either heavily or lightly treated group according to the amount of blood received when treated for hemophilia. Our goal was to investigate the possible effects of

treatment and age on AIDS incubation time. We coded the covariates $x_1 = 0$ or 1 for young or old patients and $x_2 = 0$ or 1 for light or heavy treatment. For our analysis, we assumed a lognormal model for HIV infection time with $\mu = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$. As shown in Figure 1b, the lognormal model provided a reasonably good fit to the HIV infection time except that the survival from the lognormal model was slightly higher than the Kaplan-Meier curve. In the cure model (2) for incubation time, the cure parameter $\pi = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$ and $F(t) = abt^{a-1}$. The proportion of long-term survivors (cure) was calculated as $\exp(-\pi)$.

The parameter estimates are shown in Table 3. We see that there is no significant difference in HIV infection time between different age and treatment groups. The heavily-treated group has a significantly lower probability of being AIDS-free (cure) at the end of the study. For the younger patients, the AIDS-free probabilities are 80.6% and 63.7% for the lightly and heavily-treated groups, respectively. Because $\beta_2 = 0.741$ with 95% CI (0.077, 1.436), AIDS-free probabilities are significantly different for the light and heavily treated groups.

5. Discussion

In this article, we proposed a Bayesian method to fit flexible survival models for doubly-censored data. A piecewise constant hazard function is assumed for the elapsed time. This method is easy to implement and allows a flexible class of survival models. It can be seen as a simple alternative to the maximum likelihood method. As an extension, one can consider a penalized spline for a smoothed baseline hazard.

Although the lognormal model is suitable for the two examples, the goodness-of-fit of the lognormal assumption should be validated. For diagnostics on adequacy of the lognormal distribution, one can compare the plots of the lognormal survival curves to the nonparametric survival curves based on midpoint imputation. For general applications, a more flexible model, e.g., generalized Gamma or semi-parametric model, can be used. Misspecification of $H(t|X)$ will lead to biased estimates of $S(t)$, but, the impact of model specification is not clear and needs further investigation.

Acknowledgments

The author was supported in part by the Intramural Research Program of the National Institute on Aging.

Appendix

A. WinBUGS code for the dementia survival analysis

```
Covariates Z[i,1:3]=(education, ApoE, diabetes)

Z[i,4:6]=dummy variables for (AD/VD, pure VD, other dementia)

[logTa, logTb]: Logarithm of the interval dementia onset age

Te ~ [Ta, Tb]: Dementia onset age follows truncated lognormal

Td = Death age and Ts = Ts - Te is the survival after dementia
```

Model {

```

# truncated lognormal for dementia onset age
for (i in 1:n) {

  mu.e[i]<-alpha[1]+alpha[2]*z[i,1]+alpha[3]*z[i,2]+alpha[4]*z[i,3]

  logTe[i]~djl.dnorm.trunc(mu.e[i],r.e,logtc[i],100)I(logTa[i],logTb
[i])

  Te[i]<-exp(logTe[i])
}

# priors for alpha
for (i in 1:4) {

  alpha[i]~dnorm(0,0.01)
}

r.e~dgamma(0.1,0.1)

# proportional hazards model with piecewise constant hazards
for (i in 1:n) {

  Ts[i]<-Td[i]-Te[i]

  xbeta[i]<-inprod(beta[1:6],z[i,1:6])

  for (k in 1:J) {

    # indicates event-time in interval k
    d[i,k]<-D[i]*step(Ts[i]-a[k])*step(a[k+1]-Ts[i])

    # length of overlap of t[i] with interval k
    delta[i,k]<-(min(Ts[i],a[k+1])-a[k])*step(Ts[i]-a[k])

    # the proportional hazard and piecewise constant hazard lambda
[k]

    mu[i,k]<-delta[i,k]*lambda[k]*exp(xbeta[i])

    # define the loglikelihood (if mu[i,k]=0, d[i,k]*log(mu[i,k])
=0)

    phi[i,k]<-C+mu[i,k]-d[i,k]*log(mu[i,k]+equals(mu[i,k],0))
+logfact(d[i,k])

    zeros[i,k]~dpois(phi[i,k])

  }

}

# priors for beta
for (i in 1:6) {

  beta[i]~dnorm(0,0.01)
}

```



```

}

# priors for baseline hazards of survival after dementia
for (k in 1:J) {

  lambda[k]~dgamma(1,1)

}

}

```

B. Data and WinBUGS code for the AIDS analysis

The AIDS data was extracted from

Table 1

. Biometrics 55, 909-914

n : number of individuals

(L,R): HIV infection interval

treat and group are covariates for treatment and age group

(alpha, r.e): parameters for HIV infection time

(beta, delta, lambda): parameters for AIDS incubation time

```

Model {
  for (i in 1:n) {

    # HIV infection time
    mu.e[i]<-alpha[1]+alpha[2]*group[i]+alpha[3]*treat[i]
    Te[i]~dlnorm(mu.e[i],r.e)I(ta[i],tb[i])
    Ts[i]<-Td[i]-Te[i]

    # likelihood for cure model
    s[i] <- exp(-exp(delta)*pow(Ts[i],lambda))
    xb[i]<-beta[1]+beta[2]*group[i]+beta[3]*treat[i]
    theta[i]<-exp(xb[i])
    L1[i]<- -theta[i]*(1-s[i])
    L2[i]<- D[i]*(xb[i]+log(lambda)+delta+(lambda-1)*log(Ts[i]) -exp
(delta)*pow(Ts[i],lambda))
    mu[i]<- -(L1[i]+L2[i])+C
    zeros[i] ~ dpois(mu[i])
  }
}

```

```

}
for (i in 1:3) {

  alpha[i]~dnorm(0,0.001)

  beta[i]~dnorm(0,0.001)

}

delta~dnorm(0,0.01)

lambda~dgamma(0.1,0.1)

r.e~dgamma(0.01,0.01)

b<-exp(delta)

}

```

References

1. Cai T, Cheng S. Semiparametric regression analysis for doubly censored data. *Biometrika* 2004;91:277–290.
2. Chapman DP, William SM, Strine TW, Anda RF, Moore MJ. Dementia and its implications for public health. *Preventive Chronic Disease* 2006;3(2):A34, 1–13.
3. Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. *Statistical Science* 1992;7:457–511.
4. Geweke, J. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In: Bernardo, JM.; Berger, JO.; Dawid, AP.; Smith, AFM., editors. *Bayesian Statistics*. Vol. 4. Clarendon Press; Oxford, UK: 1992.
5. Goggins WB, Finkelstein DM, Zaslavsky AM. Applying the Cox proportional hazards model for analysis of latency data with interval censoring. *Statistics in Medicine* 1999;18:2737–2747. [PubMed: 10521863]
6. Heidelberger P, Welch PD. Simulation run length control in the presence of an initial transient. *Operations Research* 1983;31:1109–1144.
7. 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]
8. Kim MY, DeGruttola VG, Lagakos SW. Analyzing doubly censored data with covariates with application to AIDS. *Biometrics* 1993;49:13–22. [PubMed: 8513098]
9. Komarek A, Lesaffre E, Harkanan T, Declerck D, Virtanen JI. A Bayesian analysis of multivariate doubly-interval-censored dental data. *Biostatistics* 2005;6:145–155. [PubMed: 15618533]
10. Kneib T. Mixed model-based inference in geoadditive hazard regression for interval-censored survival times. *Computational Statistics and Data Analysis* 2006;51:777–792.
11. Larson EB, Shadlen M, Wang L, McCormick WC, Bowen JD, Teri L, Kukull WA. Survival after initial diagnosis of Alzheimer Disease. *Annals of Internal Medicine* 2004;140:501–511. [PubMed: 15068977]
12. Law CG, Brookmeyer R. Effects of mid-point imputation on the analysis of doubly censored data. *Statistics in Medicine* 1992;11:1569–1578. [PubMed: 1439361]
13. Lui KJ, Darrow WW, Rutherford GW. A model-based estimate of the mean incubation period for AIDS in homosexual men. *Science* 1988;240:1333–1335. [PubMed: 3163848]
14. Pan W. A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics* 2001;57:1245–1250. [PubMed: 11764266]
15. Slioter AJC, Houwing-Duistermaat JJ, van Harskamp F, Cruts M, van Broeckhoven C, Breteler MB, Hofman A, Stijnen T, van Duijn CM. Apolipoprotein E genotype and progression of Alzheimer's disease: the Rotterdam Study. *Journal of Neurology* 1999;246:304–308. [PubMed: 10367700]

16. Struthers CA, Farewell VT. A mixture model for time to AIDS data with left truncation and an uncertain origin. *Biometrika* 1989;76:814–817.
17. Sun J. Empirical estimation of a distribution function with truncated and doubly interval-censored data and its application to AIDS studies. *Biometrics* 1995;51:1096–1104. [PubMed: 7548693]
18. Sun J, Liao Q, Pagano M. Regression analysis of doubly censored failure time data with applications to AIDS studies. *Biometrics* 1999;55:909–914. [PubMed: 11315027]
19. Sun L, Kim YJ, Sun J. Regression analysis of doubly censored failure time data using the additive hazards model. *Biometrics* 2004;60:637–643. [PubMed: 15339285]
20. 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* 1994;6:45–58. [PubMed: 8054493]
21. Turnbull BW. The empirical distribution function with arbitrarily grouped censored and truncated data. *Journal of the Royal Statistical Society* 1976;38:290–295.B
22. Waring SC, Doody RS, Pavlik VN, Massman PJ, Chan W. Survival among patients with dementia from a large multi-ethnic population. *Alzheimer Disease and Associated Disorders* 2005;19:178–183. [PubMed: 16327343]
23. White L, Petrovitch H, Ross GW, Masaki KH, Abbott RD, Teng EL, Rodriguez BL, Blanchette PL, Havlik RJ, Wergowske G, Chiu D, Foley DJ, Murdaugh C, Curb JD. Prevalence of dementia in older Japanese-American men in Hawaii: The Honolulu-Asia Aging Study. *Journal of the American Medical Association* 1996;276:955–960. [PubMed: 8805729]
24. Wolfson C, Wolfson DB, Asgharian M, M'Lan CE, Ostbye T, Rockwood K, Hogan DB. A reevaluation of the duration of survival after the onset of dementia. *The New England Journal of Medicine* 2001;344:1111–1116. [PubMed: 11297701]
25. Xie, J.; Brayne, C.; Matthews, FE., et al. Survival times in people with dementia: analysis from population biased cohort study with 14 year follow-up. *BMJ*. Dec 8. 2008 online from bmj.com on
26. Zhang W, Chaloner K, Cowles MK, Zhang Y, Stapleton JT. A Bayesian analysis of doubly censored data using a hierarchical Cox model. *Statistics in Medicine* 2008;27:529–542.

a. Dementia onset age

b. HIV infection time

Figure 1. Comparison of the survival function of originating event time from the lognormal model and the Kaplan-Meier method

Figure 2.
Cumulative hazards functions of survival after dementia onset

Table 1

Bias, standard error (SE) and coverage rate (CR) of the 95% CIs for parameter β_1

		Whether $H(Ix)$ depends on x					
		No			Yes		
Censoring	Method	Bias	SE	CR	Bias	SE	CR
Moderate	Exact	0.001	0.149	95.3	-0.001	0.152	94.4
	Midpoint	0.019	0.151	93.8	0.023	0.152	94.3
	SP	0.005	0.160	95.0	0.079	0.162	93.6
	Bayesian	-0.001	0.160	95.6	-0.004	0.165	94.5
Heavy	Exact	0.003	0.149	95.2	-0.002	0.152	94.7
	Midpoint	0.049	0.152	93.3	0.055	0.152	92.7
	SP	0.005	0.165	94.7	0.125	0.170	92.9
	Bayesian	-0.008	0.175	94.7	-0.009	0.176	94.6

Table 2

Parameter estimates of the survival models for dementia onset and survival after dementia

Parameter	Covariate	Estimate	SE	95% CI
Model for dementia incidence age				
α_1	Intercept	1.940	0.159	(1.621, 2.244)
α_2	Education	0.006	0.013	(-0.020, 0.032)
α_3	ApoE $\epsilon 4$	-0.208	0.111	(-0.429, 0.009)
α_4	Diabetes	-0.028	0.105	(-0.239, 0.173)
Model for survival time after dementia diagnosis				
β_1	Education	-0.017	0.022	(-0.059, 0.026)
β_2	ApoE $\epsilon 4$	-0.295	0.196	(-0.693, 0.077)
β_3	Diabetes	0.264	0.210	(-0.178, 0.655)
β_4	AD/VD	0.041	0.222	(-0.387, 0.458)
β_5	Pure VD	0.516	0.198	(0.138, 0.910)
β_6	Other	-0.053	0.206	(-0.453, 0.333)

Table 3

Parameter estimates, standard error (SE) and the 95% credible intervals (CI) for AIDS incubation time

Model for HIV infection time				
Parameter	Covariate	Estimate	SE	95% CI
α_0	Intercept	2.378	0.038	(2.302, 2.452)
α_1	Age group	0.069	0.048	(-0.025, 0.162)
α_2	Treatment	0.007	0.045	(-0.080, 0.096)
Model for AIDS incubation time				
β_0	Intercept	-1.536	0.393	(-2.222, -0.701)
β_1	Age group	0.048	0.359	(-0.673, 0.713)
β_2	Treatment	0.741	0.345	(0.077, 1.436)
a	Shape	3.315	0.605	(2.168, 4.536)
$\log(b)$	Scale	-7.572	1.168	(-10.03, -5.484)