Journal of Data Science **8**(2010), 151-172

Nonparametric Multiple Imputation of Left Censored Event Times in Analysis of Follow-up Data

Juha Karvanen, Olli Saarela and Kari Kuulasmaa
National Public Health Institute

Abstract: In this paper, we consider analysis of follow-up data where each event time is either right censored, observed, left censored or left truncated. In the case of left censoring, the covariates measured at baseline are considered as missing. The work is motivated by data from the MORGAM Project, which explores the association between cardiovascular diseases and their classic and genetic risk factors. We propose a nonparametric multiple imputation (NPMI) approach where the left censored event times and the missing covariates are imputed in hot deck manner. The left truncation due to deaths prior to baseline is compensated by Lexis diagram imputation introduced in the paper. After imputation, the standard estimation methods for right censored survival data can be directly applied. The performance of the proposed imputation approach is studied with simulated and real world data. The results suggest that the NPMI is a flexible and reliable approach to the analysis of left and right censored data.

Key words: Coronary heart disease, doubly censored data, left truncated data, MORGAM Project, proportional hazards model, survival analysis.

1. Introduction

We consider data from a follow-up study where a group of subjects (hereafter, a cohort) has been followed up to a fixed calendar period for fatal and non-fatal cardiovascular events starting from the baseline examination. Our objective is to model the effect of some covariates on the risk of coronary heart disease (CHD). The time of the first event is recorded for each subject using the age of subject as the time scale. If the first event is non-fatal, the follow-up for death continues also after the event. If no events have been occurred by the end of the follow-up, the event time is considered as right censored. If an event has occurred before the baseline examination, the event time is considered as left censored. Thus, there are three possibilities for each subject in the cohort (observed data):

1. there are no events neither during the follow-up nor before the follow-up (right censoring).

2. an event occurs during the follow-up and there has been no events before the beginning of the follow-up (event time is observed).

3. there has been an event before the beginning of the follow-up but we do not know the time when the event took place (left censoring).

Left censoring arises because it is recorded at baseline if there have been CHD events in the past but the data on the event times are not available. The event time X is observed only if it is smaller than or equal to the censoring time C and greater than B, the age of the subject at baseline. Consequently, the observed time T is obtained as a function of X, C and B

$$T = \min(\max(X, B), C).$$

The observed data can be divided into four sets: right censored observations R_0, non-fatal events during the follow-up R_1, fatal events during the follow-up R_2 and left censored observations of non-fatal events R_3. Let the numbers of the observations in these sets be n_0, n_1, n_2 and n_3, respectively.

The analysis of left censored observations requires that we change the follow-up to start from an age prior to baseline examination. This leads to the problem that, by definition, the cohort cannot have members with a fatal event before the baseline examination, and therefore the cohort followed up e.g. from age 25 is not comparable with the cohort followed up after the baseline examination. In other words, we are dealing with left truncation. Fortunately, it turns out that we can use the observed data to compensate for the potential deaths before the baseline examination. The left truncated observations can be divided into three groups. Set R_4 contains subjects who had a fatal CHD event before the baseline examination as their first event. Set R_5 contains subjects who first had a non-fatal CHD event and then later died before the baseline examination. Set R_6 contains subjects who died before the baseline examination without any preceding CHD events. Both the sets of left truncated subjects R_4, R_5 and R_6 and their numbers, n_4, n_5 and n_6, respectively, are completely unobserved.

The covariates measured at baseline examination may be divided into two categories: Sex and genes are examples of permanent covariates that do not change as a function of time. Cholesterol level, blood pressure, smoking and body mass index (BMI) are covariates that do change in time although they are often treated as constant in cohort studies. In particular, they are likely to change substantially after a CHD event due to intervention to prevent recurrent events. Therefore the values of the covariates measured after the event are influenced by the event and cannot be considered as risk factors for this event. Consequently, if the event time is left censored, the time-varying covariates must be taken as missing. We use $\mathbf{G}$ to refer to permanent covariates and $\mathbf{Z}$ to refer to time-varying covariates.

Under the assumption of non-informative censoring, the log-likelihood function related to left truncated and left and right censored data may be expressed in a general form as

$$\sum_{i\in R_0}\log(1-F(t_i|\mathbf{g}_i,\mathbf{z}_i))+\sum_{i\in R_1\cup R_2}\log(f(t_i|\mathbf{g}_i,\mathbf{z}_i))+\sum_{i\in R_3}\log(F(t_i|\mathbf{g}_i,\mathbf{Z}_i))+$$
$$\sum_{i\in R_4}\log(f(X_i|\mathbf{G}_i,\mathbf{Z}_i))+\sum_{i\in R_5}\log(f(X_i|\mathbf{G}_i,\mathbf{Z}_i))+\sum_{i\in R_6}\log(1-F(X_i|\mathbf{G}_i,\mathbf{Z}_i)),$$

where $\mathbf{z}_i$ and $\mathbf{g}_i$ refer to observed covariates, $\mathbf{Z}_i$ and $\mathbf{G}_i$ refer to unobserved covariates, and $F(t)$ and $f(t)$ are the cumulative distribution function (cdf) and the probability density function (pdf) of the event time, respectively. The different subject groups are summarized in Table 1. Our primary interest in this paper is to estimate the effect of the permanent covariates on the event times without bias and as accurately as possible. It is assumed in this paper that the reader is familiar with the standard methods for the analysis of right censored survival data.

Table 1: Different types of observations. The variables in the table are: time of event x_i, age at baseline b_i, censoring time c_i and time of death d_i.

Right censored (observed)	$R_0=\{i\,\vert\,x_i>c_i\}$
Non-fatal CHD during the follow-up (observed)	$R_1=\{i\,\vert\,b_i<x_i<c_i,\,d_i>x_i\}$
Fatal CHD during the follow-up (observed)	$R_2=\{i\,\vert\,b_i<x_i=d_i=c_i\}$
Left censored (partially observed)	$R_3=\{i\,\vert\,x_i<b_i<d_i\}$
Left truncated fatal CHD (unobserved)	$R_4=\{i\,\vert\,x_i=d_i<b_i\}$
Left truncated non-fatal CHD (unobserved)	$R_5=\{i\,\vert\,x_i<d_i<b_i\}$
Left truncated other death (unobserved)	$R_6=\{i\,\vert\,d_i<b_i,\,d_i<x_i\}$
Death during the follow-up (observed)	$R_D=\{i\,\vert\,b_i<d_i=c_i\}$
Other death during the follow-up without preceding non-fatal CHD (observed)	$R_{D0}=\{i\,\vert\,b_i<d_i=c_i<x_i\}$
Death during the follow-up with preceding non-fatal CHD (observed or partially observed)	$R_{D13}=\{i\,\vert\,b_i<d_i=c_i,\,x_i<d_i\}$

The described setup is motivated by data from the MORGAM Project (Evans *et al.*, 2005). MORGAM is a large international project on cardiovascular epidemiology that pools follow-up data from several cohorts. Currently 22 centers

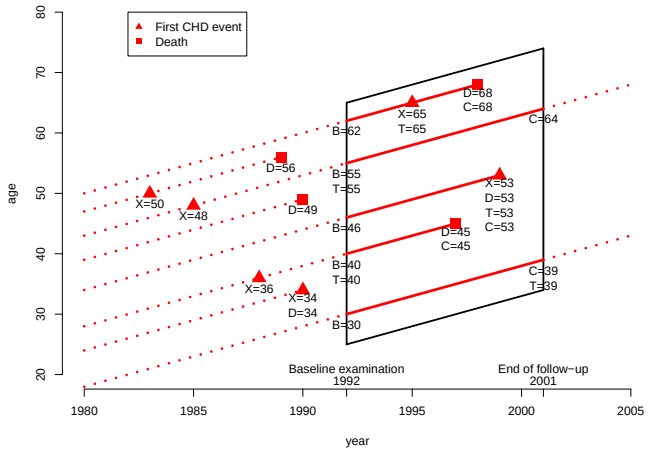

Figure 1: Illustration of a study design leading to left and right censored data with left truncation. The Lexis diagram of a cohort study is displayed. The follow-up period is from the year 1992 to the year 2001 and the age of the subjects is 25–65 years at the baseline examination. The following variables are presented: B = age at baseline examination, X = time of first CHD event, C = censoring time, T = observed time and D = time of death. In the diagram, the data of eight subjects are presented. Two subjects have an event observed during the follow-up ($X = 65$ and $X = 53$). One of the events is fatal ($X = 53$ and $D = 53$) and the other is non-fatal ($X = 65$ and $D = 68$). One subject is right censored ($C = 39$). Two subjects have a left censored event ($X = 48$ and $X = 36$). At the baseline examination, the existence of a left censored event is recorded but the exact time of an event remains unknown. One of the subjects with left censored event dies during the follow-up period ($D = 45$); the other survives up to the end of follow-up ($C = 64$). Three subjects are completely unobserved ($D = 56$, $D = 49$ and $D = 34$). One of them had fatal CHD event ($X = 34$ and $D = 34$), one had a non-fatal event ($X = 50$) and died later ($D = 56$) and one died ($D = 49$) without a preceding CHD event.

(mostly from Europe) are involved and the pooled database contains more than 140 000 subjects. The objective of the MORGAM Project is to explore the association between cardiovascular diseases (CVD) and their classic and genetic risk factors. Population cohorts, examined at study baseline, are followed up for fatal and non-fatal CVD events. The first occurrence of CHD is one of the main endpoints of the study. The MORGAM cohorts contain also subjects who had their first non-fatal CHD event before the baseline examination. For these subjects the exact event times are unknown. Although in some cases it was possible to find the exact event times e.g. from the hospital records, the cost of the additional data collection would be rather high. The percentage of subjects with baseline

CHD in MORGAM cohorts varies from 0.5 % to 13 %, which is a considerable proportion when compared to the percentage of first incidence of CHD during the follow-up that varies from 0.5 % to 17 %. Hence, the inclusion of the baseline CHD cases in the time-to-event analysis would significantly increase the number of events and provide more information on the relatively young subjects. The use of the baseline CHD cases suits well for the analysis of genetic risk factors because genotypes, contrary to many other risk factors, cannot be affected by a preceding CHD event. An illustration of a typical study design is presented in Figure 1.

In this paper, a nonparametric multiple imputation (NPMI) approach is proposed to handle the left censored and left truncated event times. In the NPMI approach, each left censored observation is replaced by several imputations drawn from empirical distribution of observed non-fatal event times. Missing covariates are imputed together with the event time. The use of the Bayesian bootstrap weights guarantees the sufficient variation between imputations. The left truncated observations are imputed as well using a novel approach that we call Lexis diagram imputation. After imputation, the standard estimation methods for right censored survival data can be directly applied. The NPMI approach follows the general idea of the multiple imputation introduced by Rubin (1987) but the essential difference is that the left censored observations are partially observed. The proposed approach can be characterized as multiple hot-deck imputation (Levy, 1998) where the set of donors is conditional on the left censored event time.

Several authors have used multiple imputation in survival analysis and public health studies. Multiple imputation of interval censored data is studied in (Pan, 2000; Glynn and Rosner, 2004; Geskus, 2001; Pan, 2001). Additional examples on the use of multiple imputation can found in (Zhou *et al.*, 2001; Taylor *et al.*, 2002; Mishra and Dobson, 2004). The main difference between these works and the approach proposed in this paper is that the NPMI does not use a parametric model for the imputation. We also impute observations that are completely unobserved.

Instead of imputation, we could, at least in principle, construct a parametric model for left and right censored data and estimate the model parameters using Bayesian approach or EM-algorithm. Nevertheless, it is not self-evident how left truncation should be taken into account in these models. If we forget the left truncation, the described setup can be seen as a special case of interval censored data where the observed intervals have form $[0, t]$ (left censored), $[t, t]$ (observed event) or $[t, \infty]$ (right censored). Examples on the analysis of interval censored data can be found e.g. in (Kim *et al.*, 1993; Zhao *et al.*, 2005; Komarek *et al.*, 2005). Alioum and Commenges (1996) proposed a method for estimation of

proportional hazards model under censoring and truncation. The method is an extension of Turnbull's nonparametric maximum likelihood estimator (Turnbull, 1974) and may suffer from identifiability problems. Our motivation to propose an imputation based solution is fourfold: First, the NPMI provides a straightforward way to deal with left censoring, left truncation and covariates not missing at random. Second, in exploratory analysis of a great number of potential covariates, the computational speed of the NPMI is an important practical benefit. Third, the use of the NPMI is not restricted to proportional hazards model. Fourth, the NPMI works as a benchmark for more complicated models.

The paper is organized as follows. The NPMI approach is introduced in Section 2. Simulation studies comparing imputed event times and covariates with their true values are presented in Section 3. Regression estimates from the NPMI approach and from the analysis of baseline healthy subjects are compared as well. A real world example using MORGAM data is presented in Section 4. Section 5 concludes the paper.

2. Nonparametric Multiple Imputation of Left Censored Event Times

2.1 Overview

In this section, we introduce a nonparametric multiple imputation (NPMI) method for the analysis of left and right censored data. Each imputation round contains generation of Bayesian bootstrap weights, imputation of left censored event times and Lexis diagram imputation of left truncated observations. The left censored event times are imputed by several values drawn from their empirical distributions. The imputation is carried out in hot deck manner selecting the donors conditionally on age at baseline examination and estimated lifetime. The missing covariates are imputed simultaneously by the covariates of the chosen donor. In Lexis diagram imputation of left truncated observations, we first generate the number of missing subjects from the Poisson distribution and then draw a random sample from all observed deaths. The sampling weights are proportional to the unobserved time in the Lexis diagram divided by the follow-up time.

2.2 NPMI with non-fatal events only

First we consider a simplified situation where all subjects are followed up at least to the age $b_{\max} = \max_i b_i$ and all events are non-fatal. We observe the age at baseline examination $b \leq b_{\max}$ and want to impute the left censored event time X. To do this we consider all subjects who had their first event during the follow-up and before age b. Let $Q = \{i \in R_1 \,|\, x_i \leq b\}$ be the set of

these subjects. If we wish our imputation scheme to be proper, each imputation round must be started with the Bayesian bootstrap (Rubin, 1987). The Bayesian bootstrap assigns a random weight w_i for each observation in the cohort. The weights are generated by taking differences of ordered uniform random numbers. More precisely, if the sample size is n, we generate $n-1$ numbers uniformly distributed on the interval $[0,1]$, sort them, include the endpoints 0 and 1, and calculate the differences of consecutive numbers. The donor j is randomly chosen from set Q with probabilities proportional to the weights w_i. The imputation for the missing event time X will be x_j and the covariates $\mathbf{z}$ are replaced by the covariates $\mathbf{z}_j$. The same weights are used for all left censored observations in one imputation round. The imputation transforms the data into right censored survival data and the standard analysis methods for such data are applicable. Imputation modifies the follow-up period to start from the age $b_{\min} = \min_i b_i$ for all subjects. After each imputation round, we fit a survival model to imputed data without the bootstrap weights and store the estimates $\hat{\boldsymbol{\beta}}_k$, where k is the number of the imputation round. The number of imputation rounds can be as small as $K = 5$ but moderate values such as $K = 20$ are preferable if we are also interested in estimating variance of $\hat{\boldsymbol{\beta}}$ reliably. After all imputation rounds, estimates $\hat{\boldsymbol{\beta}}_1, \hat{\boldsymbol{\beta}}_2, \ldots, \hat{\boldsymbol{\beta}}_K$ are combined

$$\hat{\boldsymbol{\beta}} = \frac{1}{K}\sum_{k=1}^{K}\hat{\boldsymbol{\beta}}_k \tag{2.1}$$

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}}) = \frac{1}{K}\sum_{k=1}^{K}\widehat{Var}(\hat{\boldsymbol{\beta}}_k) + \frac{K+1}{K(K-1)}\sum_{k=1}^{K}(\hat{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}})^2, \tag{2.2}$$

where the combined variance is the sum of within-imputation variance and between-imputation variance. The formulae for the combined mean and variance are the same as those routinely used for multiple imputation of missing data (Rubin, 1987).

The imputation need to be stratified by all relevant permanent covariates. This is done simply performing the imputation independently for each subgroup defined by strata. Stratification by continuous covariates requires that they are suitably categorized. If the sample size is small, it is necessary to keep the number of subgroups small in order to guarantee that there are eligible donors in each subgroup.

The use of hot deck imputation implies that the imputed event times cannot be smaller than the smallest observed non-fatal event time $x_{\min} = \min x_i$, $i \in R_1$ in the data. Therefore, successful imputation requires that the observed event times include also young subjects. If the data contains left censored observations with $t < x_{\min}$, we recommend that they are excluded from the analysis. A large

number of such observations in a data set would indicate that the NPMI is not an appropriate analysis method for the data set.

2.3 NPMI with fatal and non-fatal events

Next we consider a more realistic situation where some events are fatal and the follow-up times are shorter implying that some subjects are not followed up to the age $b_{\max}$. The imputation procedure for left censored observations is essentially the same as in Section 2.2 but the expected remaining life times must be estimated for the subjects that are withdrawn alive from the study. In addition to the notation defined above, we use D to indicate the time of death. Our cohort is sampled from the population that is alive at baseline, i.e. it always holds $D \geq T$. The right censoring has now two possible reasons: death, when $D = C$, or the end of follow-up for any other reason, when $D > C$. For each subject, age at baseline b_i, censoring time c_i and the time of death d_i (if it is not after c_i) are recorded in addition to observed time t_i and type of event. If a subject was withdrawn alive from the study, the time of death is not known but the vital status at the end of follow-up is still known. For the estimation, we need a working assumption that given the age at censoring, the time from the first event does not have an impact on the remaining life time. Statistically, $D - C$ and $C - X$ are assumed to be independent given C. The data can be used to estimate the probabilities to live an additional year on the condition that an event has occurred in the past:

$$\hat{P}(t, t+1) \equiv \hat{P}(D > t+1 \,|\, D > t, X < t) = \frac{\sum_{i \in R_1 \cup R_3} I(c_i > t+1) I(t_i < t)}{\sum_{i \in R_1 \cup R_3} I(c_i > t+1) I(t_i < t) + \sum_{i \in R_1 \cup R_3} I(t < d_i < t+1) I(t_i < t)}. \tag{2.3}$$

Estimated probabilities for survival of v additional years are obtained by chain calculations

$$\hat{P}(t, t+v) \equiv \hat{P}(D > t+v \,|\, D > t, X < t) = \hat{P}(t, t+1)\hat{P}(t+1, t+2) \ldots \hat{P}(t+v-1, t+v). \tag{2.4}$$

The imputation of left censored observations is carried out similarly to the simplified situation but the sampling probabilities for selecting the donor j are proportional to the product of the weight w_i and the estimated survival probability $\hat{P}(\lfloor c_i \rfloor, \lfloor b \rfloor)$, where the notation $\lfloor \cdot \rfloor$ refers to the full years of the age.

2.4 Lexis diagram imputation

When analyzing real world cohorts we also have to take into account that there are subjects who are excluded from the cohort due to a fatal CHD event or death for other reason prior to baseline examination and are therefore completely unknown (left truncation). In this paper we consider a novel approach where the missing subjects are imputed to the data. The approach can be motivated by a Lexis diagram (Keiding, 1998) and is therefore called Lexis diagram imputation. An illustration of the idea can be seen in Figure 2. The left panel of Figure 2 presents deaths that are observed during follow-up and deaths that are not observed because they occurred before the baseline examination. The three types of unobserved deaths correspond to the sets R_4, R_5 and R_6, and the three types of the observed deaths correspond to the sets R_2, R_{D13} and R_{D0}. T he right panel of Figure 2 shows the geometry of the Lexis diagram. We observe the deaths inside the follow-up parallelogram `ABCD` and use them to impute the deaths occurred in the triangle `ADE`. In other words, we create a cohort where the follow-up for deaths starts from the age of 25 years for everyone and subjects who did not survive until the actual baseline in the reality are represented by the imputed subjects. In the right panel of Figure 2 we have an observed death in year 1995 at age $d_i = 50$, which is represented by the point on the line segment `PG`. For this age, the line segment `PG` represents the 10 years of follow-up time and the line segment `PF` = `PA` represents $65 - 50 = 15$ years of unobserved time. Assuming that the death rates have not changed in calendar time, the expected number of the unobserved deaths corresponding the observed deaths is the ratio of `PF` to `PG` which equals to $15/10 = 1.5$. The estimated expected number of deaths in the triangle `ADE` is the sum these ratios over all observed deaths and the number of deaths to be imputed follows the Poisson distribution with this sum as the mean parameter.

The subjects who died during the follow-up are donors in the Lexis diagram imputation. The imputed subject receives the age at death as well as the type of death from the donor. Taking the Bayesian bootstrap weights into account, the weights of donors in the Lexis diagram imputation become

$$\eta_i = w_i \frac{b_{\max} - d_i}{l}, \quad i \in R_D,\ d_i \leq b_{\max}$$

where R_D is the set of subjects who died during the follow-up and l is the length of the follow-up period. The number of deaths to be imputed N_{456} is generated from the Poisson distribution with the mean parameter $N \sum \eta_i$. Then N_{456} imputations are drawn from R_D using sampling probabilities proportional to η_i. Note that this is equivalent to performing the imputation separately for R_4, R_5 and R_6.

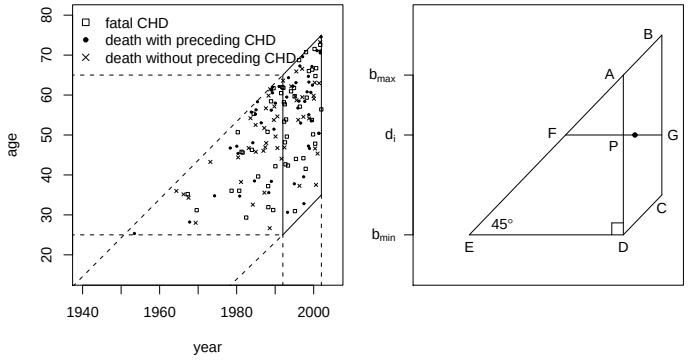

Figure 2: An illustration of Lexis diagram imputation. The left panel presents deaths that are observed during follow-up and deaths that are not observed because they occurred before the baseline examination in 1992. The right panel shows the geometry of the Lexis diagram. We are interested in all deaths inside the polygon `BCDE` but observe only the deaths inside the parallelogram `ABCD`. The deaths in the triangle `ADE` are imputed using the observed deaths. The ratio of line segments `PF` and `PG` gives the expected number of unobserved deaths corresponding to a death at age d_i.

2.5 NPMI procedure

After imputation the log-likelihood function of the data may be presented as follows

$$\sum_{i\in R_0} \log(1 - F(t_i|\mathbf{g}_i, \mathbf{z}_i)) + \sum_{i\in R_1\cup R_2} \log(f(t_i|\mathbf{g}_i, \mathbf{z}_i)) + \sum_{i\in R_3} \log(f(\hat{x}_i|\mathbf{g}_i, \hat{\mathbf{z}}_i)) +$$
$$\sum_{i\in R_4} \log(f(\hat{x}_i|\hat{\mathbf{g}}_i, \hat{\mathbf{z}}_i)) + \sum_{i\in R_5} \log(f(\hat{x}_i|\hat{\mathbf{g}}_i, \hat{\mathbf{z}}_i)) + \sum_{i\in R_6} \log(1 - F(\hat{t}_i|\hat{\mathbf{g}}_i, \hat{\mathbf{z}}_i)),$$

where $\hat{x}_i$ stands for the imputed event times and $\hat{\mathbf{g}}_i$ and $\hat{\mathbf{z}}_i$ stand for imputed permanent and time-varying covariates, respectively. Note that the choice of the survival model is independent from the imputation.

The whole procedure of the NPMI has the following steps

1. Identify the covariates used as strata and perform imputation independently for each subgroup.

2. Calculate the survival probabilities as in equations 2.3 and 2.4.

3. At the beginning of each imputation round, generate weights w_i for all observations according to the Bayesian bootstrap.

4. Sample a donor for each subject with left censored event time.

5. Apply Lexis diagram imputation to draw a random sample compensating for the left truncation.

6. Fit a survival model to imputed data without the bootstrap weights and store the estimates.

7. After a suitable number of imputation rounds, combine the estimates using equations 2.1 and 2.2.

3. Simulation Example

The performance of the NPMI approach is studied in simulations. We consider a cohort of size $n_0 + n_1 + n_2 + n_3 + n_4 = 3000$. The age at baseline B_i is generated from Uniform$(30, 65)$ distribution. The event times X_i are generated from Weibull distribution with shape parameter 8. An event is fatal with probability 0.3 and non-fatal otherwise. There are no competing causes of death. The mean event time for zero covariates is set to be $m = 65$ or $m = 80$ years and the length of the follow-up is $l = 5$ or $l = 10$ years. One thousand simulation runs are generated for each combination of mean event time and length of follow-up. Covariates Z_1 and Z_2 follow bivariate normal distribution with correlation 0.55. Covariate G is Bernoulli distributed with probability 0.5 and independent from Z_1 and Z_2. The model coefficients for Z_1, Z_2 and G are $\beta_1 = 0.2$, $\beta_2 = 0.5$ and $\alpha = 0.8$, respectively.

For the modeling we use the Cox's proportional hazard model (Cox, 1972)

$$\lambda_i(t \,|\, z_{1i},\, z_{2i},\, g_i) = \lambda_0(t) \exp(\beta_1 z_{1i} + \beta_2 z_{2i} + \alpha g_i), \tag{3.1}$$

where $\lambda(t)$ is the hazard rate, z_{1i}, z_{2i}, and g_i denotes the covariates of the ith subject and β_1, β_2 and α represent the model coefficients to be estimated. Comparisons are made between the NPMI with 20 imputation rounds and the exclusion (Excl.) approach where the left censored observations are ignored and the regression model is estimated from the rest of the data that is right-censored data. Besides the estimated model parameters, we also compare the imputed and the true distribution of left censored event times and the imputed and the true values of covariates.

The distribution of the imputed event times is studied in Figure 3. It can be seen that the empirical cumulative distribution function (cdf) of imputed event times closely resembles the empirical cdf of true left censored event times. The small difference of the curves in smallest event times can be explained by the exclusion of subjects when the left censored event time is smaller than the smallest event time during the follow-up, i.e. $b_i < x_{\min}$, $i \in R_3$. The number

of these subjects was small, five or less in 95% of simulation runs. Instead of exclusion of these subjects we also tried the use of measured covariates and age at baseline examination as event time but with exclusion the results were better. It can be also seen from Figure 3 that the distribution of event times during the follow-up clearly differs from the distribution of left censored event times.

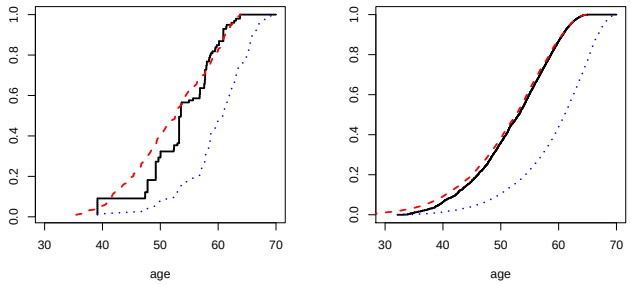

Figure 3: Imputed and true distribution of left censored event times in the simulation example. The left panel presents empirical cdfs in a typical realization and the right panel shows empirical cdfs calculated from 100 simulation runs. Solid line represents the NPMI and dashed line represents the true event times. For comparison, event times observed during the follow-up are also plotted (dotted line). The results are from Simulation B; the results from the other simulations are essentially similar.

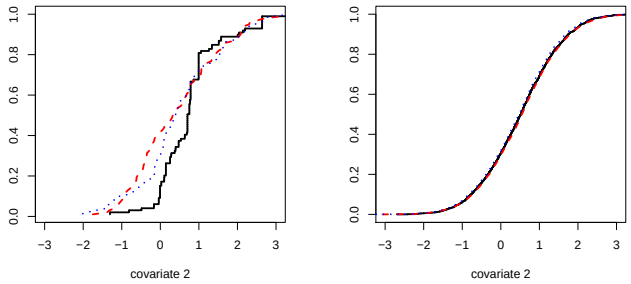

Figure 4: Imputed and true distribution of covariate z_2 for subjects with left censored event time. The left panel presents empirical cdfs in a typical realization and the right panel shows empirical cdfs calculated from 100 simulation runs. Solid line represents the NPMI and dashed line represents the true covariate values. For comparison, covariates of subjects with an event during the follow-up are also plotted (dotted line). The results are from Simulation B; the results from the other simulations are essentially similar.

Table 2: Results of the simulation example. Means of the estimates are reported together with RMSE 3.2 and the square roots of the average variances estimated from the model. The estimation methods in the comparison are the NPMI without compensation for left truncation (NPMI-N), the NPMI with the Lexis diagram imputation (NPMI-LX) and exclusion of left censored observations (Excl.). The sample size is 3000 (including left truncated subjects) and the reported numbers are means from 2000 experiments. Simulation parameter m is the mean event time for zero covariates and parameter l is the length of the follow-up. Numbers n_1, n_2 and n_3 indicate the mean number of non-fatal events, fatal events and left censored events, respectively.

	Simulation A: $m = 80$, $l = 10$ $n_1 = 186$, $n_2 = 80$, $n_3 = 89$			Simulation B: $m = 80$, $l = 5$ $n_1 = 74$, $n_2 = 31$, $n_3 = 89$		
	NPMI-N	NPMI-LX	Excl.	NPMI-N	NPMI-LX	Excl.
$\beta_1 = 0.2$	0.2019	0.2005	0.2012	0.2012	0.1994	0.1996
$\sqrt{\mathrm{Var}(\hat{\beta}_1)}$	0.0680	0.0680	0.0649	0.1120	0.1099	0.1037
$\mathrm{RMSE}(\hat{\beta}_1)$	0.0705	0.0701	0.0652	0.1275	0.1203	0.1048
$\beta_2 = 0.5$	0.5065	0.5029	0.5035	0.5116	0.5053	0.5037
$\sqrt{\mathrm{Var}(\hat{\beta}_2)}$	0.0695	0.0698	0.0666	0.1142	0.1112	0.1058
$\mathrm{RMSE}(\hat{\beta}_2)$	0.0727	0.0727	0.0665	0.1262	0.1191	0.1054
$\alpha = 0.8$	0.8135	0.8091	0.8063	0.8240	0.8172	0.8091
$\sqrt{\mathrm{Var}(\hat{\alpha})}$	0.1129	0.1187	0.1296	0.1555	0.1747	0.2093
$\mathrm{RMSE}(\hat{\alpha})$	0.1132	0.1195	0.1269	0.1643	0.1819	0.2117

	Simulation C: $m = 65$, $l = 10$ $n_1 = 451$, $n_2 = 193$, $n_3 = 352$			Simulation D: $m = 65$, $l = 5$ $n_1 = 205$, $n_2 = 88$, $n_3 = 353$		
	NPMI-N	NPMI-LX	Excl.	NPMI-N	NPMI-LX	Excl.
$\beta_1 = 0.2$	0.2062	0.2007	0.2007	0.2087	0.2011	0.2005
$\sqrt{\mathrm{Var}(\hat{\beta}_1)}$	0.0429	0.0421	0.0420	0.0663	0.0622	0.0623
$\mathrm{RMSE}(\hat{\beta}_1)$	0.0456	0.0437	0.0428	0.0702	0.0655	0.0624
$\beta_2 = 0.5$	0.5140	0.5001	0.5016	0.5199	0.5004	0.5010
$\sqrt{\mathrm{Var}(\hat{\beta}_2)}$	0.0453	0.0442	0.0442	0.0692	0.0652	0.0653
$\mathrm{RMSE}(\hat{\beta}_2)$	0.0476	0.0444	0.0437	0.0755	0.0687	0.0664
$\alpha = 0.8$	0.8212	0.8006	0.7981	0.8353	0.8061	0.8011
$\sqrt{\mathrm{Var}(\hat{\alpha})}$	0.0674	0.0708	0.0818	0.0859	0.0955	0.1214
$\mathrm{RMSE}(\hat{\alpha})$	0.0704	0.0708	0.0812	0.0941	0.0989	0.1193

The distribution of an imputed covariate is studied in Figure 4. The distributions are clearly similar and the distribution of covariates of subjects with an event during the follow-up is only slightly different. We conclude from Figures 3 and 4 that the imputed event times and covariates are unbiased or almost unbiased.

Estimated parameters of model 3.1 are presented in Table 2 for different simulation settings. Two versions of the NPMI are present: the NPMI-N does not compensate for left truncation whereas the NPMI-LX uses Lexis diagram imputation for left truncated observations. It can be seen that the NPMI-N produced biased estimates in all simulation experiments as expected. For the

NPMI-LX and the Excl. it seems that there is some bias when the number of events is small (Simulation B) but the bias decreases when the number of events increases and in Simulation C both the NPMI-LX and the Excl. produced unbiased estimates. We conclude that left truncation may have a significant effect on the estimates and recommend using the NPMI only with compensation for left truncation.

The accuracy of the estimators was measured by root mean square errors (RMSE)

$$\mathrm{RMSE}(\hat{\beta}_{\mathrm{method}}) = \sqrt{\mathrm{Mean}\left((\beta_{\mathrm{true}} - \hat{\beta}_{\mathrm{method}})^2\right)}, \tag{3.2}$$

where $\hat{\beta}_{\mathrm{method}}$ is one of following $\hat{\beta}_{\text{NPMI-N}}$, $\hat{\beta}_{\text{NPMI-LX}}$, $\hat{\beta}_{\text{Excl}}$. The RMSEs are compared to the square root of the average of variance estimates from 2000 simulation runs. It can be seen that the square roots of the estimated variances are close to RMSEs. Comparison between the NPMI and the Excl. reveals that the Excl. resulted slightly smaller RMSEs for parameters β_1 and β_2 but for parameter α the NPMI was clearly better in terms of RMSEs. It also seems that the difference in RMSEs for parameters β_1 and β_2 between the NPMI and the Excl. decreases when the number of events increases. In simulation C, the RMSEs of β_1 and β_2 were almost the same for the NPMI and the Excl. The result can be understood if we consider the amount of information available in the NPMI and the Excl. Covariate G is observed for all subjects and consequently exclusion of observations in the Excl. increases the variance of parameter α. On the other hand, covariates Z_1 and Z_2 are missing for left censored subjects and the imputed covariates contain same amount information as the non-imputed covariates. A small number of events handicaps the NPMI but when the number of events is sufficiently large the variances of covariates Z_1 and Z_2 should be the same for the NPMI and the Excl.

4. Example with Real Data

To test the NPMI with real world data we consider FINRISK cohorts (Kulathinal *et al.*, 2005; Vartiainen *et al.*, 2000) that are a part of the MORGAM Project. The baseline examinations of the cohorts were in 1982, 1987, 1992 and 1997, and all cohorts were followed up to the end of year 2001 except 1997 cohorts that were followed up to the end of year 2003. In our example, an event is defined as the first incidence of the CHD (fatal or non-fatal) and the follow-up is set to end at the age of 75 years at the latest. The numbers of different events for each cohort are summarized in Table 3. The age at baseline varies from 25 years to 64 years (the upper limit was 74 years in 1997 cohorts but we exclude subjects over 64 years at baseline). The NPMI is suited for this data because there are

practically no CHD events before the age of 25 years.

Table 3: Summary of the FINRISK cohorts used in the example. n is the total number of subjects, n_1 and n_2 indicate the number of non-fatal and fatal events during the follow-up and n_3 is the number of left censored events.

Cohorts	Men				Women			
	n	n_1	n_2	n_3	n	n_1	n_2	n_3
1982	4465	478	209	124	4564	248	85	45
1987	2802	176	80	132	3009	99	26	54
1992	2833	126	30	120	3166	51	8	31
1997	3293	77	18	110	3640	22	8	38
Combined	13393	857	337	486	14379	420	127	168

Cox's proportional hazard model explaining the event times by classic CHD risk factors is fitted to the data. Age of the subject is used as the time scale. The covariates in the model are the ratio of total to high-density lipoprotein (HDL) cholesterol (MORGAM variable `RCHOL`), mean of systolic and diastolic blood pressure (`BPM`), body mass index (`BMI`), daily smoking (yes or no, `DSMOKER`) and family history of CHD (defined as the answer to the question: "Has your father had any of the following diseases before the age of 60 years: myocardial infarction or angina pectoris", `FHISCHD`). The model is fitted separately for men and women and is stratified by the cohort baseline year and the geographical region (East or West).

We compare two approaches for the left censored event times: the NPMI and the Excl. In the Excl. approach 486 men and 168 women are removed from the analysis due to CHD event prior to baseline. Additional 187 men and 127 women are removed because of stroke prior to baseline and 7 men and 6 women are removed because missing covariate measurements (mainly missing `BMI`). Further, 15 men and 5 women are excluded because of very high `RCHOL` values (> 15). High values of the cholesterol ratio indicate that total cholesterol is very high and/or HDL cholesterol is very low. Both very high total cholesterol and very low HDL cholesterol are associated with high CHD risk but are best handled as special cases. We are interested in modeling the risk of `RCHOL` values widely represented in our cohorts and outlying values may have an undesirable impact on the parameter estimates. Missing values of covariates `DSMOKER` and `FHISCHD` are combined with category 'no'.

In the NPMI, left censored event times and covariates `RCHOL`, `BPM`, `BMI` and `DSMOKER` are imputed. Imputation is stratified by sex, family history (`FHISCHD`), the year of baseline and the region. Family history (`FHISCHD`) is taken as a permanent covariate although in principle there is a chance for a change from 'no' to 'yes' as time passes. The 187 men and 127 women who had stroke but

not myocardial infarction prior to baseline were excluded from the imputation and the analysis. Same exclusion criteria for missing covariates and for very high `RCHOL` values as in the Excl. approach was used.

The number of imputation rounds is 20. The sunflower plot in Figure 5 presents five imputed event times for each left censored observation. There are relatively few observed events for younger subjects which causes the same donor to be used in several times and is seen as overlapping points in the plot. Figure 5 also illustrates the difference in the CHD incidence between men and women.

Figure 5: Imputed event times in FINRISK. The number of sunflower leaves is equal to the number of multiple observations.

The distribution of the imputed covariates for $R_3 \cup R_4 \cup R_5$ is shown in Figure 6. The distribution of the covariates of the subjects with an event during the follow-up is plotted for comparison. For men the distributions are rather similar but

for women there differences especially in BMI. The difference or equality of the distributions does not tell about the performance of the imputation but reflects the changes in the covariates as a function of age.

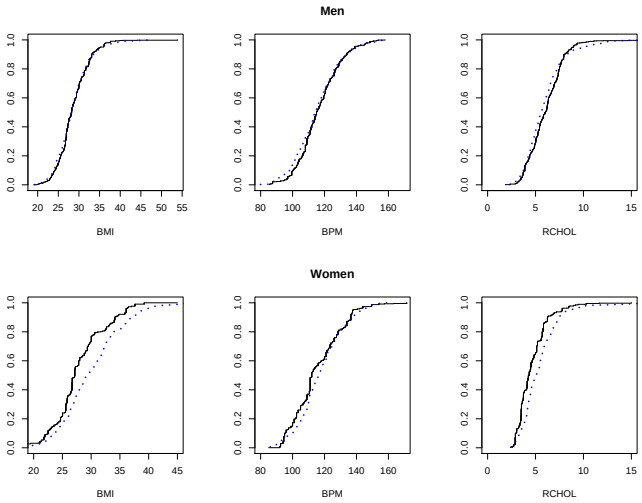

Figure 6: Distribution of imputed covariates (solid line) and distribution of covariates of subjects with an event during the follow-up (dotted line) in FINRISK.

The parameter estimates from the Cox's proportional hazard model are summarized in Table 4. Covariates `BPM`, `RCHOL`, `DSMOKER` and `FHISCHD` have a statistically significant effect at 95 % risk level in all models. `BMI` is significant for men. These results are in the agreement with the previous knowledge of these association. The variance of the NPMI estimate of the permanent covariate `FHISCHD` is smaller than the variance of the Excl. estimate as we could except on the basis of the simulation example. The variance of the other covariate are in general slightly smaller for the Excl. estimates, which is also in the agreement with the simulation results. There are some differences between the NPMI estimates and the Excl. estimates that might not be completely explicable by random variation. We do not have a comprehensive explanation for the differences but the potential explanations include non-proportional hazards, changes in the covariates in the time and unbalanced cohort due to non-response. The detailed analysis of these data will be carried out as a part of the general MORGAM analysis plan.

Figure 7 illustrates the relative importance of the covariates in the cohort. Each covariate is ordered in ascending order and the relative hazard compared to the median of the covariate is plotted as a function of cumulative covariate distribution. This gives insight on the epidemiological significance of the covariates and makes it possible to compare covariates measured on the different scale.

Table 4: Estimated covariate effects for the FINRISK data. Estimates and their standard errors (se) are presented together with relative hazard estimates (exp(estim.)) and their 95 % confidence intervals.

Men: Without baseline CVD (Excl.)				
	estimate	se	exp(estim.)	conf. int.
BPM (mmHg)	0.0165	0.0022	1.0167	(1.0123,1.0210)
RCHOL	0.2071	0.0159	1.2301	(1.1924,1.2690)
BMI (kg/m^2)	0.0166	0.0082	1.0168	(1.0005,1.0333)
DSMOKER (0/1)	0.5312	0.0624	1.7010	(1.5052,1.9222)
FHISCHD (0/1)	0.3610	0.0659	1.4348	(1.2611,1.6325)

Men: Imputed event times and covariates for baseline CHD cases (NPMI)				
	estimate	se	exp(estim.)	conf. int.
BPM (mmHg)	0.0063	0.0026	1.0063	(1.0013,1.0114)
RCHOL	0.2034	0.0158	1.2255	(1.1881,1.2642)
BMI (kg/m^2)	0.0270	0.0086	1.0274	(1.0102,1.0448)
DSMOKER (0/1)	0.6167	0.0663	1.8528	(1.6269,2.1101)
FHISCHD (0/1)	0.4725	0.0572	1.6040	(1.4339,1.7942)

Women: Without baseline CVD (Excl.)				
	estimate	se	exp(estim.)	conf. int.
BPM (mmHg)	0.0156	0.0033	1.0158	(1.0093,1.0223)
RCHOL	0.2967	0.0257	1.3454	(1.2793,1.4149)
BMI (kg/m^2)	0.0141	0.0099	1.0142	(0.9948,1.0340)
DSMOKER (0/1)	0.5723	0.1308	1.7724	(1.3715,2.2904)
FHISCHD (0/1)	0.3140	0.1013	1.3689	(1.1224,1.6695)

Women: Imputed event times and covariates for baseline CHD cases (NPMI)				
	estimate	se	exp(estim.)	conf. int.
BPM (mmHg)	0.0106	0.0032	1.0106	(1.0043,1.0170)
RCHOL	0.2558	0.0273	1.2914	(1.2242,1.3624)
BMI (kg/m^2)	0.0027	0.0105	1.0027	(0.9822,1.0236)
DSMOKER (0/1)	0.6602	0.1332	1.9353	(1.4904,2.5128)
FHISCHD (0/1)	0.4303	0.0890	1.5378	(1.2917,1.8307)

According to Figure 7, `RCHOL` and `DSMOKER` seem to be the most serious covariates in the FINRISK cohorts. The differences between the NPMI and the Excl. estimates of `DSMOKER`, `FHISCHD` and `BPM` and are visible also in Figure 7.

5. Conclusion

In this paper, we considered the estimation of regression models from left and right censored survival data and proposed the NPMI approach that converts the left and right censored data into multiple right censored data sets. The left truncation due to deaths prior to baseline is compensated by Lexis diagram imputation. The imputation of left censored data is done without reference to the underlying distribution or model of the event time and hence the procedure can be applied to more general model than the Cox's proportional hazards model. In simulations it was found that the distributions of the imputed event times and

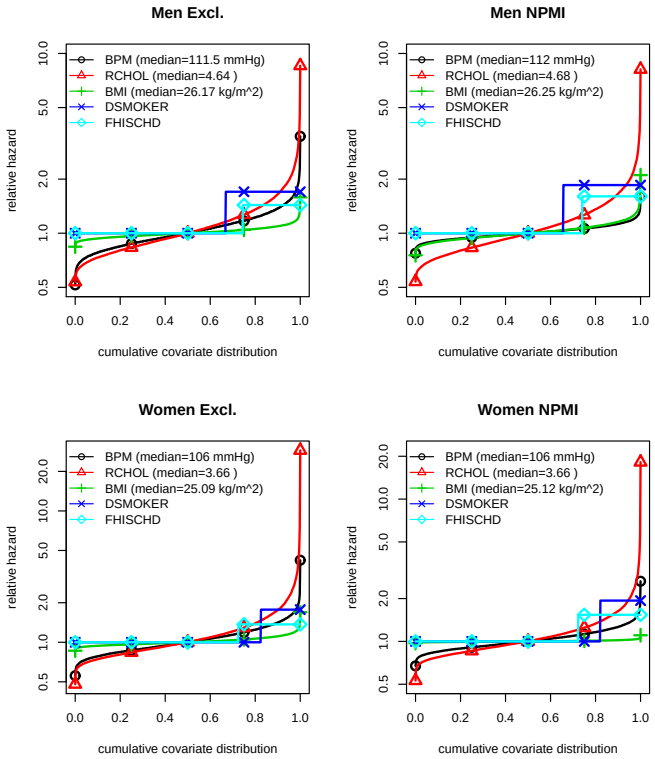

Figure 7: Relative hazard as a function of ordered covariates. Relative hazard is calculated respect to the median of the covariate that is displayed in the legend. Cumulative probability is used in the x-axis instead of the actual covariate values to allow displaying all covariates in the same plot.

the imputed covariates are very close to the true distributions. Good performance was also observed when analyzing real world data from FINRISK cohorts.

The NPMI is specifically designed for the data arising from the MORGAM Project but the approach may be applicable for other studies as well. The main requirement for the NPMI approach is the existence of prospective follow-up data for all relevant ages. In the genetic sub-study of MORGAM, the NPMI need to be adapted to the case-cohort design. This does not require any changes to imputation itself.

Compared to the Excl. the main benefit of the NPMI is the gain of efficacy when estimating the effect of permanent covariates. This has practical importance in the MORGAM Project where one of the main goals is the testing of candidate genes. The primary interest is then on the statistical significance of the candidate genes and the classic risk factors in the model have secondary importance. Compared to parametric imputation, the NPMI is robust against to

imputation model misspecification. Compared to full likelihood alternatives (e.g. EM algorithm or Bayesian methods) the benefits of the NPMI are speed and straightforward implementation (standard methods and software may be used). In fact, we are not aware of any practical full likelihood based approach that would be directly applicable to the described setup with left truncation. The drawbacks of the NPMI are the need of multiple analyses and small loss of efficacy compared to the Excl. when estimating the effect of imputed covariates from data with small number of the events. A small bias might be also unavoidable if the number of the events is small.

The results in this paper suggest that the inclusion of left censored observations without compensating for left truncation leads to biased estimates. This conclusion is not restricted to the NPMI but applies to all analyses where the events may be fatal or non-fatal and the follow-up is modified to start from a time prior to the recruitment of the cohort. The bias, however, is not necessarily large compared to the standard errors of estimates in moderately sized cohort studies.

Acknowledgments

This work was supported by the GenomEUtwin Project grant from the European Commission under the programme Quality of Life and Management of the Living Resource of 5th Framework Programme (no. QLG2-CT-2002-01254) and by the Academy of Finland via its grant number 53646. The authors would like to thank Dr. Sangita Kulathinal for helpful discussions and the FINRISK study group for the data used in the examples.

References

Alioum, A. and Commenges, D. (1996). A proportional hazards model for arbitrarily censored and truncated data. *Biometrics* **52**, 512-524.

Regression models and life-tables (with discussion). *Journal of Royal Statistical Society Series B* **34**, 187-220.

Evans, A., Salomaa, V., Kulathinal, S., Asplund, K., Cambien, F., Ferrario, M., Perola, M., Peltonen, L., Shields, D., Tunstall-Pedoe, H., and Kuulasmaa for the MORGAM Project, K. (2005). Cohort profile: MORGAM (an international pooling of cardiovascular cohorts). *International Journal of Epidemiology* **34**(1), 21-27.

Geskus, R. B. (2001). Methods for estimating the AIDS incubation time distribution when date of seroconversion is censored. *Statistics in Medicine* **20**, 795-812.

Glynn, R. J. and Rosner, B. (2004). Multiple imputation to estimate the association between eyes in disease progression with interval-censored data. *Statistics in Medicine* **23**, 3307-3318.

Keiding, N. (1998). Lexis diagram. In *Encyclopedia of Biostatistics*. New York: Wiley.

Kim, M. Y., DeGruttola, V. G., and Lagakos, S. W. (1993). Analysing doubly censored data with covariates with application to AIDS. *Biometrics* **49**, 13-22.

Komarek, A., Lesaffre, E., Härkänen, T., Declerck, D., and Virtanen, J. I. (2005). A Bayesian analysis of multivariate doubly interval censored dental data. *Biostatistics* **6**(1), 145-155.

Kulathinal, S., Niemelä, M., Kuulasmaa, K., and contributors from Participating Centres for the MORGAM Project (2005). Description of MORGAM cohorts, Available from URL:http://www.ktl.fi/publications/morgam/cohorts/, URN:NBN:fife20051214.

Levy, P. S. (1998). Missing data estimation "hot deck" and "cold deck". In *Encyclopedia of Biostatistics*. New York: Wiley.

Mishra, G. D. and Dobson, A. J. (2004). Multiple imputation for body mass index: lessons from the Australian longitudial study on women's health. *Statistics in Medicine* **23**, 3077-3087.

Pan, W. (2000). Multiple imputation approach to Cox regression with interval censored data. *Biometrics* **56**, 199-203.

Pan, W. A. (2001). A multiple imputation approach to regression analysis for doubly censored data with application to AIDS studies. *Biometrics* **57**, 1245-1250.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley Series in probability and mathematical statistics. New York: John Wiley & Sons, Inc.

Taylor, J. M. G., Murray, S., and Hsu, C.-H. (2002). Survival estimation and testing via multiple imputation. *Statistics and Probability Letters* **58**, 221-232.

Turnbull, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *Journal of the American Statistical Association* **69**, 169-173.

Vartiainen, E., Jousilahti, P., Alfthan, G., Sundvall, J., Pietinen, P., and Puska, P. (2000). Cardiovascular risk factor changes in Finland, 1972-1997. *International Journal of Epidemiology* **29**, 49-56.

Zhao, X., Lim, H., and Sun, J. (2005). Estimating equation approach for regression analysis of failure time data in the presence of interval-censoring. *Journal of Statistical Planning and Inference* **129**, 145-157.

Zhou, X.-H., Eckert, G. J., and Tierney, W. M. (2001). Multiple imputation in public health research. *Statistics in Medicine* **20**, 1541-1549.

Received November 30, 2007; accepted March 26, 2008.

Juha Karvanen
International CVD Epidemiology Unit
Department of Health Promotion and Chronic Disease Prevention
National Public Health Institute
Mannerheimintie 166, 00300
Helsinki, Finland
juha.karvanen@ktl.fi

Olli Saarela
International CVD Epidemiology Unit
Department of Health Promotion and Chronic Disease Prevention
National Public Health Institute
Mannerheimintie 166, 00300
Helsinki, Finland
olli.saarela@ktl.fi

Kari Kuulasmaa
International CVD Epidemiology Unit
Department of Health Promotion and Chronic Disease Prevention
National Public Health Institute
Mannerheimintie 166, 00300
Helsinki, Finland
kari.kuulasmaa@ktl.fi