# Locally Efficient Estimation with Current Status Data and Time-Dependent Covariates

Mark J. van der Laan and James M. Robins[1]

May 15, 2001

**Abstract**

In biostatistical applications interest often focuses on the estimation of the distribution of a failure time-variable $T$. If one only observes whether or not $T$ exceeds an observed monitoring time $C$, then the data structure is called current status data, also known as interval censored data, case I. We extend the data structure by allowing the presence of a possibly time-dependent covariate process which is observed up till the monitoring time $C$. We follow the approach of Robins and Rotnitzky (1992) by modeling the hazard of $C$ conditional on the failure time-variable and the covariate-process, i.e. the missingness or censoring process, under the restriction that the missingness (monitoring) process satisfies coarsening at random.

Because of the curse of dimensionality no globally efficient nonparametric estimators with

[1]Mark J. van der Laan is Assistant Professor, School of Public Health, Division of Biostatistics, University of California, Berkeley, CA 94720. James M. Robins is Professor of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA 02115. The authors thank two anonymous referees and the associate-editor for their helpful remarks and Alan Hubbard for his work on the simulations and data-analysis. The authors thank Steve Shiboski and Nancy Padian for providing the data-set. This research was supported by a FIRST award from the National Institute of General Medical Sciences, NIH (GM53722), and a NIH-grant R01-AI32475.

a good practical performance at moderate sample sizes exist. We introduce an inverse probability of censoring weighted estimator of the distribution of $T$ and of smooth functionals of this distribution of $T$ which are guaranteed to be consistent and asymptotically normal if we have available a correctly specified parametric or semiparametric model for the missingness process. Furthermore, given a correctly specified model for the missingness process, we propose a locally efficient one-step estimator whose asymptotic variance attains the efficiency bound, if we can correctly specify a lower-dimensional model for the conditional distribution of $T$ given the covariates. The estimator remains consistent and asymptotically normal even if this latter submodel is misspecified. We conclude with a simulation experiment and a data analysis.

KEY WORDS: Current status data, Asymptotically linear estimator, Influence curve, Asymptotically efficient, Cox proportional hazards model.

# 1 Introduction.

Consider a study in which interest lies in the distribution $F$ of a failure time random variable $T$. Suppose $T$ is never observed. Rather, for each individual, we observe at a random monitoring (censoring) time $C$ whether $T$ exceeds $C$. This data structure is called *current status data*.

Previous work and examples of current status data can be found in Diamond, McDonald and Shah (1986), Jewell and Shiboski (1990), Diamond and McDonald (1991), Keiding (1991), Sun and Kalbfleisch (1993), among several others. In its nonparametric setting it is also known as interval censoring, case I (Groeneboom and Wellner, 1992). Current status data commonly arise in epidemiological investigations of the natural history of disease and in animal tumorigenicity experiments. Jewell, Malani and Vittinghoff (1994) give two examples that arise from studies of Human Immunodeficiency Virus (HIV) disease.

In each of the aforementioned papers, inferences on the distribution of $T$ were made under the assumptions that (i) $T$ and $C$ are independently distributed and (ii) no data are available on additional time-independent or time-dependent covariates. Under (i) and (ii), the nonparametric maximum likelihood estimator (NPMLE) of $F$ is the pooled adjacent violators estimator for the estimation of the monotone regression $F(t) = E(\Delta \mid C = t)$ of Barlow, Bartholomew, Brenner, and Brunk (1972), where $\Delta = I(T \leq C)$ is the current status indicator at time $C$. The asymptotic distribution of this estimator under (i) has been analyzed by Groeneboom and Wellner (1992). Under assumptions (i)-(ii), the efficiency of the NPMLE of smooth functionals of $F$ (such as its mean and variance) has been proved by Groeneboom and Wellner (1992), van de Geer (1994), and Huang and Wellner (1995). The purpose of this paper is to develop methods for the estimation of smooth functionals of $F$ when (ii) and/or (i) is violated.

It will be pedagogically advantageous to motivate our methods by considering the following idealized study design for a mouse tumorigenicity experiment in which the outcome of interest is time to the development of liver adenoma and study mice are randomly allocated to either placebo or active treatment with a suspected tumorigen. Suppose liver adenomas are never, in themselves, the primary cause of an animal's death. Therefore, each mouse is sacrificed (monitored) at a random time $C$; at autopsy it is determined whether a tumor has developed before $C$. In such studies, it is easy to collect daily measurements of the weight of each mouse prior to sacrifice. Let $L(u)$ be the weight at time $u$ and let $L = L(\cdot)$ be the entire weight process. We only observe the weight process up to time $C$, i.e. we observe $\bar{L}(C) = \{L(u) : 0 < u < C\}$. Since mice with liver adenomas tend to lose weight, $\bar{L}(C)$ and $T$ are correlated. Suppose, for the moment, that the sacrifice times $C$ are randomly chosen by the investigators, guaranteeing independence of $T$ and $C$. A major goal of this paper is to develop estimators of smooth functionals of $F$ that, by incorporating information on the surrogate marker process $\bar{L}(C)$, are guaranteed both (a) to be more efficient than the NPMLE that ignores data on $\bar{L}(C)$, and (b) to remain consistent and asymptotically normal, whatever be the joint distribution of $(T, L)$. Specifically, we will propose a new class of "inverse probability of censoring weighted" (IPCW) estimators that accomplish both goals (a) and (b). In contract to our weighted estimators, the NPMLE that incorporates data on $\bar{L}(C)$ fails to attain these goals because of the curse of dimensionality (Robins and Ritov, 1997); indeed, the NPMLE is not even well-defined.

The results described in the previous paragraph required the assumption that the monitoring time $C$ was assigned completely at random. However, in realistic settings, this will often not be so, in which case the assumption that $T$ and $C$ are independent will also be false. As a consequence, the NPMLE that ignores data on $L$ will be inconsistent both for $F$ and for smooth functionals of $F$. As an example, suppose in our mouse tumorigenicity experiment, the investigators wish, if possible, to sacrifice an animal soon after it has developed a tumor in order to obtain more efficient estimates of $F$. To do so, they decide to increase the probability of sacrificing an animal in the interval $(t, t + \delta t)$ if the animal has begun to lose weight prior to $t$. Under this design, $T$ and $C$ will be dependent (since they both are correlated with the weight process) but the hazard of censoring (sacrifice) at $t$ given the weight process $\bar{L}(t)$ up to $t$ will not further depend on the unknown failure time $T$. That is,

$$\lambda(t \mid T, L) = \lambda(t \mid \bar{L}(t)), \tag{1}$$

where $\lambda(t \mid T, L)$ is the hazard of censoring, given the full data $(T, L)$. Formally, assumption (1) says that the conditional hazard of $C$ at time $t$, given the full data $T, L$, is only a function of $(t, \bar{L}(t))$. We will show that when $T$ and $C$ are dependent but (1) is true, our inverse probability of censoring weighted estimators of smooth functionals of $T$ will remain consistent and asymptotically normal provided that we can consistently estimate the hazard function $\lambda(t \mid \bar{L}(t))$. In the idealized experiment described above, $\lambda(t \mid \bar{L}(t))$ will be known by design since it is under the control of the investigator. However, in many settings, $\lambda(t \mid \bar{L}(t))$ will not be known. For example, in our tumorigenicity experiment, it may happen that a number of animals die of other non-tumor related diseases prior to sacrifice, at which point they are autopsied and the presence or absence of a liver adenoma determined. In such animals, their monitoring time is their time of death $C$. For that case one cannot be sure that (1) holds. As a consequence, the investigator should try to incorporate in $\bar{L}(t)$ any factors that could be correlated with the presence of an underlying adenoma (i.e. correlated with $T$) that also predict the hazard of death (censoring) at time $t$. Even were the investigator successful so that (1) were true, nonetheless the functional form of $\lambda(t \mid \bar{L}(t))$ will remain unknown. In this setting it is necessary to posit a statistical model for $\lambda(t \mid \bar{L}(t))$. In this paper, we will emphasize modeling $\lambda(t \mid \bar{L}(t))$ by a time-dependent Cox-proportional hazards model. Our IPCW estimators will be consistent if (1) holds and our Cox proportional hazards model for $\lambda(t \mid \bar{L}(t))$ is correctly specified.

In current practice carcinogenicity experiments have few predetermined times of sacrifice (see e.g. Dinse, 1988) and do not collect time-dependent measurements (e.g. weight) till sacrifice. In this case, our estimators can still be used to produce efficient estimates of $F(\cdot)$ at the sacrifice times. However, note that such carcinogenicity experiments have an inferior design relative to our idealized design described above since they 1) do not allow nonparametric estimation of smooth functionals of $F$ 2) use an inefficient sacrificing scheme independent of $T$ and 3) do not exploit the possibility of using a surrogate process $\bar{L}(C)$ to improve estimation.

The critical assumption (1) is not empirically testable since $T$ is not observed. The need for untestable assumptions such as (1) and for the correct specification of models for high dimensional processes, such as $\lambda(t \mid \bar{L}(t))$, is the rule in analyzing data obtained from observational studies. Our critical assumption (1) is equivalent to the assumption that the censoring (i.e., missingness) mechanism satisfies coarsening at random (CAR) (Robins and Rotnitzky, 1992, Robins, 1996). Coarsening at random was originally formulated by Heitjan and Rubin (1991) and generalized by Jacobsen and Keiding (1994) and Gill, van der Laan, Robins (1997). The usefulness of the CAR assumption in estimation of $F$ in the presence of time-dependent surrogate processes $\bar{L}(t)$ has been argued by Robins and Rotnitzky (1992). In Gill, van der Laan, Robins (1997) it is shown that if CAR is the only assumption, then the model for the observed data is nonparametric and all regular asymptotically linear estimators of $\mu$ are asymptotically equivalent and efficient. By the curse of dimensionality this means that estimators with reasonable moderate sample performance do not exist in this unrestricted model. However, when $\lambda(t \mid \bar{L}(t))$ follows a semiparametric model, the estimators proposed in this paper will perform reasonably in moderate samples and are still efficient at a submodel of interest. It is important to stress again that when the CAR assumption (1) holds but $T$ and $C$ are dependent, no alternatives to the approach described herein are available in the literature.

The paper is organized as follows. In section 2, we formalize our estimation problem. In section 3, we introduce the class of inverse probability weighted estimators. We then use in section 4 a preliminary inverse probability of censoring weighted estimator as an initial estimator and propose a one-step update that leads to a locally efficient estimator. Our proposed locally efficient estimator requires that we specify a model for the conditional distribution of $T$ given $\bar{L}(u)$. The one-step estimator is locally efficient in the sense that, whenever $\lambda(t \mid \bar{L}(t))$ is either known or correctly modeled, it attains the semiparametric variance bound for the model if, in fact, the model for $T$ given $\bar{L}(u)$ is correct and yet remains consistent and asymptotically normal with good efficiency properties even if the latter model is misspecified. Indeed in the extreme case in which the surrogate process $L(\cdot)$ is time-independent and happens to have correlation 1 with $T$ (which is unknown to the experimenter) and the selected model for $T$ given $L$ is correct, then, remarkably, the inability to observe $T$ results in no loss of information. That is our proposed locally efficient update of the IPCW estimator is asymptotically equivalent to the estimator obtained from the empirical distribution of $T_1, \ldots, T_n$. Similarly, if $\bar{L}(t)$ happens to have correlation 1 with $I(T \leq t)$, e.g. in the above example all mice first begin to loose weight immediately upon onset of the tumor, then our proposed one-step estimator is asymptotically equivalent with the Kaplan-Meier estimator based on $(T_i \wedge C_i, I(T_i \leq C_i))$, $i = 1, \ldots, n$, if the selected model for $F(t \mid \bar{L}(u))$ is correct. The consistency of our one-step estimator under misspecification of the model for $T$, given $\bar{L}(u)$ is protected by our assumption of a correctly specified model for the hazard $\lambda(t \mid T, L) = \lambda(t \mid \bar{L}(t))$ of censoring given $(T, L)$. The only price we pay for this robustness is that the calculation and estimation of the asymptotic variance of our one-step estimator requires explicit computation of projections on various Hilbert spaces analogous to those in Robins (1993, 1996) and Robins and Rotnitzky (1992) for right-censored data.

In section 5, we present the local efficiency theorem and a consistent estimator for the asymptotic variance of our locally efficient one-step estimator which we then use to construct Wald confidence intervals for the functional of interest. The proof of the theorem is deferred to our technical report van der Laan, Robins (1997). In section 6, we compare our methods to the NPMLE for the marginal model in a simulation experiment. In section 7, we use our methods to reanalyze data on time to HIV infection in the female partners of HIV infected males in the California Partners Study. We conclude with a discussion. Here we address, in particular, how one can generalize our methodology to estimate the onset distribution at a given point and to handle discrete monitoring times.

## 2 A formalization of our problem.

For simplicity, we will assume that $T$ has support on a finite interval $[0, \tau]$. Let $L = \{L(u) : 0 < u < \tau\}$. For each individual we observe $Y = (C, \Delta = I(T \leq C), \bar{L}(C))$. Given a real valued function $r$, we are concerned with nonparametric estimation of functionals $\mu(r) \equiv \int r(t)(1 - F(t))dt$ of the distribution of $F$ of $T$, based on $n$ i.i.d. observations $Y_i = (C_i, \Delta_i, \bar{L}(C_i))$. We now show that $\mu(r)$ includes the mean and all higher order moments of $T$ for appropriate choices of $r(\cdot)$. Let $R(x) = \int_0^x r(t)dt$. Notice that by integration by parts we have:

$$\int r(t)(1 - F(t))dt = R(t)(1 - F(t)) \mid_0^\tau + \int R(t)dF(t).$$

Hence if $\lim_{t \to \tau}(1 - F(t))R(t)$ is zero, then an estimate of $\int r(1 - F)dt$ provides us with an estimate of $\int RdF$. In particular, if $r(t) = 1$, then $\mu(r) = ET$ and if $r(t) = kt^{k-1}$, then $\mu(r) = ET^k$. Moreover, by setting $r(t) = K(\{t - t_0\}/h)/h$ for some kernel $K$ and bandwidth $h$ an estimator of $\mu(r)$ provides us with an estimator of $S(t_0) = 1 - F(t_0)$.

For notational convenience, we shall also use $L$ to denote the function $L(\cdot)$. This same shorthand will be used for other functions. In addition, define $X \equiv (T, L)$ and let $F_X$ be the distribution of $X$. We will make no assumptions about this distribution. Let $G(\cdot \mid X)$ be the conditional distribution function of $C$ given $X$; it will also be referred to as the censoring mechanism or missingness process. It will be assumed that $G(\cdot \mid X = x)$ is absolutely continuous w.r.t. the Lebesgue measure with density $g(c \mid x)$ which satisfies:

$$g(c \mid (T, L)) = h(c, \bar{L}(c)) \text{ for some function } h \text{ of } (c, \bar{L}(c)). \tag{2}$$

Let $\lambda(c \mid x = (t, l))$ be the hazard corresponding with $g(c \mid x)$: $\lambda(c \mid x) = g(c \mid x)/\bar{G}(c \mid x)$ with $\bar{G}(c \mid x) = 1 - G(c \mid x)$. Because $\bar{l}(c)$ determine $\bar{l}(u)$, $u \leq c$, we have that $g(c \mid x) = \lambda(c \mid x) \exp\left(-\int_0^c \lambda(u \mid x)du\right)$ is only a function of $\bar{l}(c)$ if and only if

$$\lambda(c \mid X = (t, l)) = m(c, \bar{l}(c)) \text{ for some function } m \text{ of } (c, \bar{l}(c)). \tag{3}$$

In turn, equation (3) is equivalent to equation (1) of the introduction. Condition (2) implies that the censoring mechanism satisfies *coarsening at random* (CAR) (i.e. it is non-informative, given the observed covariates). In the appendix of van der Laan, Robins (1997) we prove that (2) is in fact equivalent to the apparently weaker assumption of CAR. Note that if $L$ is time-independent and actually thus uncensored, i.e. $\bar{L}(C) = L$, then equation (3) states that $C$ and $T$ are independent, given the covariates $L$. A special case of CAR is that $T$ and $C$ are marginally independent so that $g(c \mid x) = g(c)$. Gill, van der Laan and Robins (1997) show that the assumption (2) that the data are CAR is the only assumption places no restrictions on the joint distribution of the observed data: that is the observed data model is completely nonparametric.

As discussed in the introduction, whenever the density $g(c \mid x)$ is not known by design, it is usually necessary to model it in order to avoid the curse of dimensionality. We assume a parametric or semiparametric model for $g(c \mid x)$ or equivalently for $\lambda(c \mid x)$,indexed by a parameter $\eta$:

$$g(c \mid x) = h_\eta(c, \bar{l}(c)), \tag{4}$$

where $h_\eta$, $\eta \in \Gamma$, is a parametric or semiparametric parametrization of $h$. In particular, we will emphasize the Cox-proportional hazards model as a model for $\lambda(c \mid x)$. Our inverse probability of censoring weighted estimators require that we obtain $n^{-1/4}$-consistent estimator of $g(c \mid x)$. Thus if one assumes a Cox proportional hazard, then the usual discrete estimator of the baseline cumulative hazard must be replaced by a smooth version, as discussed further in section 4 below. $g(c \mid x)$ can be estimated by smoothing the discrete density obtained from the partial likelihood estimators for the Cox-proportional hazards model, as done in our simulation section.

## 3 Inverse Probability of Censoring Weighted Estimators.

The key identity that we exploit to construct our IPCW estimators is the following:

$$\begin{aligned} E\left(\frac{(1-\Delta)r(C)}{g(C \mid X)}\right) &= EE\left(\frac{(1-\Delta)r(C)}{g(C \mid X)} \mid X\right) = E\left(\int_0^\tau I(T \geq c)r(c)dc\right) \\ &= \int_0^\tau r(c)(1-F(c))dc. \end{aligned}$$

This suggest the following IPCW estimator of $\mu = \int r(1-F)dt$:

$$\mu_n^0 = \frac{1}{n}\sum_{i=1}^n \frac{I(\Delta_i = 0)r(C_i)}{g_{\eta_n}(C_i \mid X_i)}, \tag{5}$$

where $g_{\eta_n}$ is a consistent estimator of $g$ obtained assuming the model $\{g_\eta : \eta \in \Gamma\}$. Recall that by (2), $g_\eta(c \mid x) = g_\eta(c \mid L = l) = h_\eta(c, \bar{l}(c))$ and hence is a function of the observed data. By setting $r(t) = K(\{t - t_0\}/h)/h$ for a kernel $K$ and bandwidth $h$ we obtain the following estimator of $S(t_0) = 1 - F(t_0)$:

$$S_n^0(t_0) = \frac{1}{n}\sum_{i=1}^n \frac{(1-\Delta_i)\frac{1}{h}K(\{C_i - t_0\}/h)}{g_{\eta_n}(C_i \mid X_i)}, \tag{6}$$

where $K$ is a kernel and $h$ a bandwidth which converges to zero at an appropriate rate.

For example, if $g(c \mid x) = g(c)$, i.e. we have independent censoring, and if $g_{\eta_n}$ is a kernel density estimator of $g$ based on the $C_1, \ldots, C_n$, then $F_n^0$ is just the regularized MLE in the current status model (without a covariate process) as studied by van der Laan, Bickel and Jewell (1997). For smooth functionals such as the moments of $F$, $\mu_n^0$ will be asymptotically linear where an estimator $\mu_n$ is asymptotically linear with influence curve $IC$ if it can be approximated by a sum of i.i.d. random variables in the following sense:

$$\mu_n - \mu = \frac{1}{n}\sum_{i=1}^n IC(C_i, \Delta_i, \bar{L}_i(C_i)) + o_P(1/\sqrt{n}),$$

for some function $IC$ of $(C, \Delta, \bar{L}(C))$ with mean zero and finite variance. Then $\sqrt{n}(\mu_n - \mu)$ is asymptotically normal with mean zero and variance given by the variance of the influence curve. Van der Laan (1995) provides conditions under which the estimator $\mu_n^0$ is asymptotically linear. The most important condition is that, for every $(T, L)$ and all $c$ in the support $[0, \tau]$ of $T$ with $r(c) > 0$, $g(c \mid T, L) > \epsilon > 0$ for some $\epsilon > 0$. Such a condition is not unexpected since $F(t) = E(\Delta \mid C = t)$ if $T$ is independent of $C$.

## 4 The locally efficient one step estimator.

In this section, we construct a locally efficient one-step estimator by adding to the estimator $\mu_n^0$ (5) an estimate of the empirical mean of the efficient influence function. Our first task therefore is to provide a representation of the efficient influence function. This representation has two pieces. The first is given by the influence function of $\mu_n^0$ when using the known $g(c \mid X)$ which is given by:

$$IC_0(C, \Delta, \bar{L}(C) \mid \mu, G) = \frac{r(C)(1-\Delta)}{g(C \mid X)} - \mu. \tag{7}$$

The second piece is a projection of $IC_0$ on the so called orthogonal complement to the tangent space defined in section 5 below. The projection is a function $IC^*_{nu}(\cdot \mid F_X, G)$ of $(C, \Delta, \bar{L}(C))$:

$$IC^*_{nu} = \int \left\{ \frac{r(u)\{1 - F(u \mid \bar{L}(u)\}}{g(u \mid X)} - \frac{1}{\bar{G}(u \mid X)} \left\{ \int_u^\infty r(t)\{1 - F(t \mid \bar{L}(u)\} dt \right\} \right\} dM(u), \tag{8}$$

where

$$dM(u) \equiv I(C \in du) - \Lambda(du \mid X) I(C > u). \tag{9}$$

It is important to emphasize that, for any function $H(u, \bar{L}(u))$ the stochastic integral

$$\int H(u, \bar{L}(u)) dM(u) = H(C, \bar{L}(C)) - \int_0^C H(u, \bar{L}(u)) \Lambda(du \mid X)$$

is the function of the observed data because $\lambda(u \mid X)$ depends on $X$ only through $\bar{L}(u)$. Similarly, the integrand in (8) only depends on $X$ through $\bar{L}(u)$.

Sometimes, as on the left-hand side of (8), we will suppress the dependence of functions on both $F_X$, $g$ and on the data. In the appendix in van der Laan, Robins (1997) we show that the efficient influence curve for estimation of $\mu$ is

$$IC^*(C, \Delta, \bar{L}(C) \mid F_X, G, \mu) \equiv IC_0(C, \Delta, \bar{L}(C) \mid \mu, G) - IC^*_{nu}(C, \Delta, \bar{L}(C) \mid F_X, G).$$

If $L$ is a vector of time-independent covariates and thus always completely observed (i.e. $\bar{L}(C) = L$), then $IC^*_{nu}$ reduces to

$$IC^*_{nu}(C, \Delta, L \mid F_X, G) = \frac{r(C)}{g(C \mid X)}\{1 - F(C \mid L)\} - \int r(u)\{1 - F(u \mid L)\}du. \tag{10}$$

Let $IC^*_{nu}(\cdot \mid F_{X,n}, G_n)$ be an estimator of $IC^*_{nu}(\cdot \mid F_X, G)$ obtained by substitution of estimators of $F(t \mid \bar{L}(u))$, $t \geq u$, and $g(c \mid x)$, where $G_n = G_{\eta_n}$ suppresses, in the notation, the dependence on the parameter $\eta$. Note that $IC^*_{nu}$ depends on $G$ through $g(c \mid X)$ and through the measure $dM(u)$. Theorem 5.1 below shows that asymptotic normality of the one-step estimator requires that the estimate of $g(c \mid X)$ in (7)-(8) must be at least $n^{-1/4}$-consistent for $g(c \mid X)$, thus requiring a smooth. Smooth estimation of $g(c \mid X)$ according to a given model is discussed in the simulation section. In contrast, theorem 5.1 also teaches us that we can estimate $\Lambda(du \mid X)$ in $dM(u)$ with a discrete estimator of $G$; so there is no need to evaluate integrals w.r.t. continuous measures. In the next subsection we will propose an estimation method for $F(t \mid \bar{L}(u))$ in (8). The locally efficient one step estimator is given by:

$$\mu_n^1 = \mu_n^0 + \frac{1}{n}\sum_{i=1}^{n}\left\{IC_0(C_i, \Delta_i, \bar{L}_i(C_i) \mid \mu_n^0, G_n) - IC^*_{nu}(C_i, \bar{L}_i(C_i) \mid F_X^n, G_n)\right\}, \tag{11}$$

where $\mu_n^0$ is the estimator (5) using $g_n(c \mid x)$. Let $Pf \equiv \int f dP$ for a probability measure $P$ and measurable function $f$. Let $P_n$ be the empirical CDF so that $P_n f = 1/n \sum_{i=1}^n f(Y_i)$. Note that, in fact, $P_n IC_0(\cdot \mid \mu_n^0, G_n) = 0$ so that we could delete the $IC_0$-term in (11), as did Robins and Rotnitzky (1992). However, we chose to include the $IC_0$-term in the representation (11) in order to show that $\mu_n^1$ is just the classical one-step estimator as defined in BKRW (page 395);

that is, by its definition, $\mu_n^1$ is the first step in the Newton-Raphson algorithm for solving the estimating equation

$$0 = \frac{1}{n} \sum_{i=1}^{n} IC^*(C_i, \Delta_i, \bar{L}_i(C_i) \mid F_X^n, G_n, \mu) \tag{12}$$

for $\mu$, where we chose $\mu_n^0$ as the initial estimator. This follows from the fact that the derivative of the estimating equation (12) with respect to $\mu$ equals $-1$.. Since (12) is linear in $\mu$, the Newton-Raphson algorithm converges at the first step so that $\mu_n^1$ equals the solution of (12). It is easy to understand that one can improve on a given estimating equation for $\mu$, say $0 = \sum_{i=1}^n IC_0(Y_i \mid \mu, G_n)$ solved by $\mu_n^0$, by replacing the equation by another estimating equation with smaller variance but the same derivative w.r.t. $\mu$. The estimating equation (12) does this by subtracting from $IC_0$ an estimate of an optimally chosen negatively correlated random variable which is not a function of $\mu$, namely $IC_{nu}^*(Y \mid F_X^n, G_n)$. This provides a heuristic explanation of the fact that $\mu_n^1$ typically improves on $\mu_n^0$.

### 4.1 A method for estimation of $IC_{nu}^*$.

Consider first the case in which $L$ is time-independent and thus fully observed. Then $IC_{nu}^*$ is given by (10) and consequently estimation of $IC_{nu}^*$ requires estimation of $F(u \mid L_i)$ for observations $i = 1, \ldots, n$ for $u$ in the set $\mathbf{C} = \{C_1, \ldots, C_n\}$ of observed monitoring times.

In a particular application a natural model for $F(u \mid L) = P(T < u \mid L)$ might arise and one could estimate $F(u \mid L)$ with the maximum likelihood procedure. Here we provide a generic estimation method which is based on an additive logistic model. Suppose that $L$ is a $k$-dimensional vector of covariates: so $L = (L_1, L_2, \ldots, L_k) \in \mathbb{R}^k$. By CAR, $T$ and $C$ are independent given $L$. Hence

$$F(c \mid L) = P(T < c \mid C = c, L) = E(\Delta \mid C = c, L), \tag{13}$$

where again $\Delta = I(T < C)$. Consequently, we can estimate $F(c \mid L)$ by estimating the regression of $\Delta$ on the $k+1$ dimensional vector $(C, L)$. To avoid the curse of dimensionality, we assume a generalized additive logistic regression model:

$$E(\Delta \mid C = c, X = (t, l)) = \frac{1}{1 + \exp(f_0(c) + f_1(l_1) + \ldots + f_k(l_k))}$$

for unknown functions $f_0, \ldots, f_k$, where $f_0$ is constrained to be monotone. The functions $f_0, \ldots, f_k$ can be estimated using the SPLUS-function "GAM". GAM allows the user to specify a parametric model for one or more of the functions $f_m$ if the analyst so chooses (Hastie and Tibshirani, 1990). The model considered by Rossini and Tsiatis (1996) is a special case of this model in which $f_1, f_2, \ldots, f_k$ are enforced to be linear. Rossini and Tsiatis (1996) consider estimation of the model using a sieve-MLE estimator. In contrast, Huang (1994) modeled the distribution of $T$ given $L$ with a Cox proportional hazards model and carried out estimation via maximum likelihood.

Consider now the case in which $L(u)$ is a time-dependent covariate process and we observe $\bar{L}(C)$. Then $IC^*_{nu}$ is given by (8). Thus it is necessary to estimate $F(t \mid \bar{L}(u))$ for a given $(t, u)$ with $t \geq u$. The estimator we propose is motivated by the following CAR-identity:

$$F(t \mid \bar{L}(u)) = E(J(u,t) \mid \bar{L}(u), C > u)$$

with

$$J(u,t) = \frac{\bar{G}(u \mid X) I(C \geq t) F(t \mid \bar{L}(t))}{\bar{G}(t \mid X)}.$$

Firstly, for fixed $t$ construct summary measures $W_1, \ldots, W_k$ (i.e. functions) of $\bar{L}(t)$, which hopefully contain the most relevant information for predicting $T$. Then we model $F(t \mid \bar{L}(t), C > t) = E(\Delta \mid C = t, \bar{L}(t), C > t)$ with an additive logistic regression model, where the functions $f_m(l_m)$ are now replaced by functions $f_m(w_m)$, and use GAM to obtain an estimator $\hat{F}(t \mid \bar{L}(t))$

using the subsample with $C_i > t$, $i = 1, \ldots, n$. This results in an estimator

$$\hat{J}(u,t) = \frac{\bar{G}_n(u \mid X) I(C \geq t) \hat{F}(t \mid \bar{L}(t))}{\bar{G}_n(t \mid X)}$$

of $J(u,t)$. Finally, $\hat{F}(t \mid \bar{L}(u))$ is obtained by regressing $\hat{J}_i(u,t)$ on functions of $\bar{L}(u)$, among subjects with $C > u$.

If the density of $C$ only depends on the time-independent covariates, i.e. $g(c \mid X) = g(c \mid L(0))$, then we can simplify the procedure above by noting that

$$F(t \mid \bar{L}(u)) = E(\Delta \mid C = t, \bar{L}(u), C > u).. \tag{14}$$

In this case we can estimate $F(t \mid \bar{L}(u))$ with GAM as above by regressing $\Delta$ on $C$ and summary measures of $\bar{L}(u)$, among the subjects with $C > u$, where one repeats this GAM-procedure for each $u \in \mathbf{C}$.

# 5 Local efficiency result and construction of asymptotic confidence intervals.

When the model for $F(t \mid \bar{L}(u))$ is misspecified, the asymptotic variance of $\mu_n^1$ will depend on the model for the nuisance parameter $g(c \mid x)$. To characterize the form of this dependence requires that we introduce the notion of a tangent space.

Denote the Hilbert space of functions of $(C, \Delta, \bar{L}(C))$ with finite variance and mean zero, endowed with the covariance norm $\|v\|_{P_{F_X,G}} \equiv \sqrt{\int v^2 dP_{F_X,G}}$, by $L_0^2(P_{F_X,G})$. The tangent space $T_1(P_{F_X,G})$ for the parameter $F_X$ is, by definition, the closure of the linear extension in $L_0^2(P_{F_X,G})$ of the scores at $P_{F_X,G}$ from correctly specified parametric models for the distribution $F_X$ of $X$. The tangent space $T_2(P_{F_X,G})$ for the parameter $G_\eta$ is the closure of the linear extension in $L_0^2(P_{F_X,G})$ of the scores at $P_{F_X,G}$ from all correctly specified parametric submodels (i.e.

submodels of the assumed semiparametric model) for the distribution $G$. For convenience, we will often denote these tangent spaces by $T_1$ and $T_2$, suppressing the dependence on $P_{F_X,G}$.

The CAR assumption (2) implies that the spaces $T_1$ and $T_2$ are mutually orthogonal. Given this fact it follows from BKRW (1993) that the efficient influence curve is given by:

$$IC^* = IC_0 - IC^*_{nu}, \tag{15}$$

where $IC^*_{nu} = \Pi(IC_0 \mid T_1^{\perp})$ and $T_1^{\perp}$ is the orthogonal complement of $T_1$. Here $\Pi(\cdot \mid T_1^{\perp})$: $L^2(P_{F_X,G}) \to L^2(P_{F_X,G})$ is the projection operator on $T_1^{\perp}$. It is proved in van der Laan, Robins (1997) that the explicit form of $IC^*_{nu}$ is given by (8).

From BKRW we have that an estimator $\mu_n$ is asymptotically efficient if it is asymptotically linear with influence curve $IC^*$. Theorem 5.1 below shows that if our model $F(t \mid \bar{L}(u))$ is correctly specified, the one-step estimator $\mu_n^1$ is indeed asymptotically linear with influence curve $IC^*$ and thus is asymptotically efficient. Moreover, $\mu_n^1$ has the additional feature that it remains a consistent and asymptotically normal estimator of $\mu$ even when our model for $F(t \mid \bar{L}(u))$ is misspecified. This is due to the fact that, by (8), $IC^*_{nu} = \int H(u, \bar{L}(u))dM(u)$ for a particular function $H$ where the stochastic integral $\int H(u, \bar{L}(u))dM(u)$ has conditional mean zero, given $X$, for any function $H$, since $E(dM(u) \mid X) = 0$. This explains why $\mu_n^1$ will still be consistent if $H$ is estimated inconsistently.

With these preliminaries, we are now ready to state our main theorem. Before doing so, we note that condition (ii) in the theorem below is a general empirical process condition. For empirical process theory we refer to van der Vaart and Wellner (1996). We decided not to derive more primitive conditions that imply condition (ii) since condition (ii) is technical and model dependent. Recall the notation $Pf = \int f(x)dP(x)$.

**Theorem 5.1** *We assume*

*(i)* $r(c)/g(c \mid l) < M < \infty$ *a.e w.r.t.* $dF_{C,L}(c, l)$*, for some* $M$*.*

*(ii)* $IC_0(\cdot \mid \mu_n^0, G_n) - IC_{nu}^*(\cdot \mid F_{X,n}, G_n)$ *falls in a* $P_{F_X,G}$*-Donsker class with probability tending to 1.*

*(iii)* $F_n(t \mid \bar{L}(u))$ *converges uniformly in* $(t, \bar{L}(u))$, $t \geq u$, *to a* $F^1(t \mid \bar{L}(u))$ *and* $g_n(c \mid L)$ *converges uniformly in* $(c, L)$ *to* $g(c \mid L)$, *both on the support of* $P_{F_X,G}$.

*(iv)*

$$\sup_{c,L} \mid g_n(c \mid L) - g(c \mid L) \mid \sup_{t \in [u,\tau], L} \mid F_n(t \mid \bar{L}(u)) - F^1(t \mid \bar{L}(u) \mid = o_P(1/\sqrt{n}), \tag{16}$$

$\tau$ *being the endpoint of the support of* $T$ *and where the supremum is taken over the support of* $P_{F_X,G}$.

*(v)* $\Phi(G_n)$ *is an asymptotically efficient estimator of* $\Phi(G)$ *for a model containing the true* $G$ *with tangent space* $T_2(P_{F_X,G})$.

*Then* $\mu_n^1$ *is asymptotically linear with influence curve given by*

$$IC(\cdot \mid F_X^1, G, \mu) \equiv \Pi\{IC_0(\cdot \mid \mu, G) - IC_{nu}^*(\cdot \mid F_X^1, G) \mid T_2^{\perp}(P_{F_X,G})\},$$

*In particular, if* $IC_{nu}^*(\cdot \mid F_X^1, G) = IC_{nu}^*(\cdot \mid F_X, G)$, *then* $\mu_n^1$ *is asymptotically efficient.*

*For the case that our semiparametric model for* $g(c \mid x)$ *is characterized by the restriction* $g(c \mid x) = g(c)$ *for an unspecified density* $g$ *we have for any* $V \in L^2(P_{F_X,G})$ $\Pi(V \mid T_2) = E(V(C, \Delta, \bar{L}(C)) \mid C)$. *For the case in which* $\bar{L}(u)$ *is possibly time-dependent and* $g(c \mid x)$ *follows a Cox proportional hazards model with* $k$*-dimensional covariate* $\vec{W}(\bar{L}(u)) = (W_1(\bar{L}(u)), \ldots, W_k(\bar{L}(u))^\top \in \mathbb{R}^k$:

$$\lambda(c \mid x) = \lambda_0(c) \exp(\beta^\top \vec{W}(\bar{L}(c))), \tag{17}$$

*we have*

$$\begin{aligned}\Pi(V \mid T_2) &= \int E(V^*(u, \bar{L}(u)) \mid C = u) dM(u) \\ +E\left(\int V^*(u, \bar{L}(u)) dM(u) \left\{\int \vec{W}'(u, \bar{L}(u)) dM(u)\right\}^\top\right) & \Sigma^{-1} \int \vec{W}'(u, \bar{L}(u)) dM(u),\end{aligned} \tag{18}$$

*where*

$$\begin{aligned}\vec{W}'(u,\bar{L}(u)) &\equiv \vec{W}(\bar{L}(u)) - E(\vec{W}(\bar{L}u) \mid C = u)\\ V^*(u,\bar{L}(u)) &\equiv E\left(V(u,I(u<T),\bar{L}(u)) - V(C,\Delta,\bar{L}(C)) \mid C > u, \bar{L}(u)\right)\end{aligned} \qquad (19)$$

*and the $(i,j)$'th element of the covariance matrix $\Sigma$ is given by:*

$$\int\int W_i'(u,\bar{L}(u))W_j'(u,\bar{L}(u))\lambda(u \mid L)Y(u)dM(u).$$

Condition (v) will typically be satisfied when $g_n(\cdot \mid X)$ is a smoothed maximum likelihood estimator, where the smooth needs to be chosen so that $g_n(\cdot \mid X)$ is at least $n^{-1/4}$-consistent. Note that theorem 5.1 also gives the limiting distribution of the estimator $\mu_n^0$; by (8) simply choose the estimator $F_n(t \mid \bar{L}(u))$ to be identically 1 for all $t$ and $u$. Then $\mu_n^1 = \mu_n^0$ and $F_X^1(t \mid \bar{L}(u))$ is also identically 1. Furthermore, note that when in truth $g(c \mid L) = g(c)$, the marginal censoring model and the Cox-proportional hazards model are both true. However, the tangent space $T_2$ associated with the less restrictive Cox proportional hazards censoring model strictly includes the tangent space $T_2$ associated with the marginal censoring model. This implies, by theorem 5.1, that the estimator $\mu_n^0$ (5) that estimates $g(c \mid X)$ by (i) imposing the Cox model, (ii) estimating the coefficient $\beta$ by partial maximum likelihood, and (iii) estimating the baseline hazard function using a kernel smooth is at least as efficient as the estimator $\mu_n^0$ that estimates $g(c \mid X)$ by ignoring data on $L$ and applying a kernel density estimator to the data $C_i$, $i = 1,\ldots,n$. As discussed earlier, this latter estimator is asymptotically equivalent to the NPMLE $\hat{\mu}_{NPMLE}$ that ignores data on $L$. Thus, when censoring is independent, i.e. $g(c \mid X) = g(c)$, we have proved that, as promised in the introduction, we are able to construct an estimator that is guaranteed to be at least as efficient as $\hat{\mu}_{NPMLE}$ that ignores data on $L$.

**Construction of Wald-type confidence interval.** Consider the case where we impose the model $g(c \mid x) = g(c)$ with $g$ unspecified and we estimate the unknown density $g(c)$ using a kernel density estimator $g_n$ that ignores data on $L$ with appropriately chosen band width. Then the influence curve $IC(\cdot \mid F_X^1, G, \mu)$ of $\mu_n^1$ can be estimated as the residual from the nonparametric estimation of the regression of $IC_0(Y_i \mid \mu_n^0, G_n) - IC_{nu}^*(Y_i \mid F_X^n, G_n)$ on $C_i$, $i = 1, \ldots, n$. The resulting estimator $\widehat{IC}$ of $IC$ is under the assumptions of theorem 5.1 $L^2(P_{F_X,G})$-consistent. Since $\sqrt{n}(\mu_n^1 - \mu)$ is asymptotically normal with mean zero and variance equal to the variance of $IC(C_i, \Delta_i, \bar{L}_i(C_i) \mid F_X^1, G, \mu)$, we can construct an asymptotic Wald confidence interval for $\mu$ by estimating the variance of $IC(C_i, \Delta_i, \bar{L}_i(C_i) \mid F_X^1, G, \mu)$ by

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{IC}^2(C_i, \Delta_i, \bar{L}_i(C_i)),$$

where $\widehat{IC}$ is consistent for $IC$. If $F_n(t \mid \bar{L}(u))$ is consistent for $F(t \mid \bar{L}(u))$, $\widehat{IC}$ will converge to the efficient influence curve. In practice, however, it will not be known whether $F_n(t \mid \bar{L}(u))$ is consistent. However, whether it is or not, the actual coverage rate of our Wald intervals will converge to its nominal rate in large samples. In the case where $g(c \mid x)$ follows the Cox-proportional hazards model (17), it is necessary to estimate $\Pi(V \mid T_2)$ as given in theorem 5.1.

## 6 Some simulation results.

In the simulations we refer to the estimator $\mu_n^0$ that estimates $g(c \mid X)$ by a kernel density estimator applied to the data $C_i$, $i = 1, \ldots, n$ as the estimator "univariate". We refer to the estimator $\mu_n^0$ based on the Cox proportional hazards model as the estimator "Cox". In the simulations below, we smoothed the usual discrete baseline hazard estimator for the Cox model using a Gaussian kernel with an edge correction and bandwidth selected by cross-validation. The consistency properties of such a kernel smoother are derived in Andersen, Borgan, Gill and Keiding (1993). Finally, we denote the one-step estimator $\mu_n^1$ as "1-step".

In this section, we compare the performance of the competing estimators "PAV", i.e. the NPMLE $\hat{\mu}_{NPMLE}$ which ignores data on $L$, "Cox" and "1-step" in a number of simulation experiments. It is useful to remember that the estimator "PAV" is only consistent for $\mu$ under the assumption that $g(c \mid X) = g(c)$. In all our simulation experiments, the functional $\mu$ was the truncated mean survival time $\int r(t)(1 - F(t))dt$, where either $r(t) = I(t \leq \tau)$ or $r(t) = I(\tau_1 < t < \tau_2)$. In all experiments we used a univariate time-independent covariate $L$ which is uniformly distributed on the interval $[-1, 1]$.

The data generating process of the first simulation experiment was

$$
\begin{aligned}
g(c \mid l) &= \frac{\exp(-10 + c)}{(1 + \exp(-10 + c))^2} \exp(1.75l) \\
F(c \mid l) &= \frac{1}{1 + \exp(-10 + c + 1.75l)}.
\end{aligned}
$$

Under this data generating mechanism, $T$ and $C$ are marginally dependent but independent given the covariate $L$. So PAV is inconsistent. Further, the density $g(c \mid l)$ in this and all subsequent simulations lies in the Cox proportional hazards model (17) with $\vec{W}(\bar{L}(u)) = L$ which was assumed in computing the estimators Cox and one-step. We set $r(t) = I(\tau_1 < t < \tau_2)$, where $\tau_1, \tau_2$ are the 0.1 and 0.9 quantile of the marginal distribution of $C$. We need to do this because we can only estimate $F$ well on an interval where $g(c \mid l) > \delta > 0$ and a full mean would use all values of $F$. In our first simulation we compare our proposed estimators of the truncated mean $\mu = \int_{\tau_1}^{\tau_2} (1 - F)(t)dt$ with the PAV estimator which ignores the covariates. We estimated $F(c \mid l)$ by fitting, by maximum likelihood, the linear logistic model

$$
\text{logit} E(\Delta \mid C = t, L = l) = \alpha_0 + \alpha_1 t + \alpha_2 l, \tag{20}
$$

which is correctly specified under our data generating process. It follows that the estimator "1-step" will be semiparametric efficient. Results are given in table (1). In terms of mean squared error, the PAV is beaten by "Cox" which in turn is outperformed by the efficient estimator

Table 1: Mean squared error (MSE) for estimation of $\mu = \int_{\tau_1}^{\tau_2}(1-F)(t)dt$, $n = 500$ based on 625 replicates. $C$ and $T$ depend on each other through $L$ and we guess the right model for $F(c \mid L)$.

| Estimator | 100× MSE | Relative MSE |
|---:|---:|---:|
| PAV | 4.3 | 1.0 |
| Cox | 3.6 | 1.2 |
| One-step | 1.2 | 3.7 |

"1-step". We remark that the gain of the IPCW-estimator relative to the PAV depends on the data generating distribution and the model chosen for the data generating mechanism.

In the next simulation study we chose

$$g(c \mid l) = f(c \mid l) = \lambda(l)\exp(-\lambda(l)c)$$

where $\lambda(l) = 0.25\exp(2l)$. Under this data generating process, we were able to obtain stable estimators by choosing $r(t) = I(t \leq \tau)$, where $\tau$ is now the 0.95 quantile of the distribution of $C$. In constructing the estimator "1-step" we again estimated $F(c \mid l)$ using the logistic model (20). Note that under this second data generating process, the logistic model (20) is misspecified so our estimator $F(c \mid l)$ is inconsistent. Results are given in table (2). Again note that both "Cox" and "1-step" outperform the inconsistent estimator "PAV". However, the estimator "1-step" no longer is more efficient than "Cox", presumably because $F(c \mid l)$ was badly biased.

In the following simulation, $C$ is distributed uniformly $[0, 10]$ and independently of $T$. Specifically,

$$\begin{aligned} g(c \mid l) &= \frac{1}{10} \\ F(c \mid l) &= \frac{1}{1+\exp(-30+6c+10l)}. \end{aligned}$$

We estimated $F(c \mid l)$ consistently with the linear logistic regression model (20). For this

Table 2: MSE for estimation of $\int_0^\tau (1 - F(t))dt$, $n = 500$ based on 625 replicates. $C$ and $T$ depend on each other through $Z$ and we guess the wrong model for $F(c \mid Z)$.

| Estimator | 100× MSE | Relative MSE |
|---:|---:|---:|
| PAV | 69.4 | 1.0 |
| Cox | 11.0 | 6.3 |
| One-step | 22.7 | 3.1 |

Table 3: MSE for estimation of $ET$, $n = 500$ based on 625 replicates. $C$ and $T$ independent and we guess the right-model for $F(c \mid l)$.

| Estimator | 100× MSE | Relative MSE |
|---:|---:|---:|
| PAV | 1.2 | 1.0 |
| Univariate | 1.2 | 0.9 |
| Cox | 1.2 | 0.9 |
| One-step (Kernel) | 0.5 | 2.1 |
| One-step (Cox) | 0.5 | 2.1 |

simulation with independent censoring we add both the "univariate" estimator (that uses a kernel density estimator which ignores data on $L$ for the estimate of $g(c)$) and a 1-step estimator that also uses the same kernel density for the estimate of $g(c)$. The results are reported in table 3. In this simulation the "Univariate", "PAV" and "Cox" have a similar performance while the "1-step" outperforms them. These results are in agreement with theoretical asymptotic relative efficiencies of the estimators in Table 3 except that asymptotically Cox is more efficient than PAV or univariate.

## 7 Data analysis.

We obtained data on 88 female partners of HIV infected males from the California Partners' Study. The California Partners' Study is an ongoing investigation of heterosexual HIV-transmission in partners of infected index cases (Padian et al., 1987, Shiboski and Jewell, 1992). Participants are recruited through referrals from physicians, research studies, and local departments of public health. Upon entry, serum samples are drawn from study subjects to determine status with regard to HIV-infection. In addition, detailed medical, contraceptive, and behavioral histories are obtained. Finally, couples are interviewed to determine the total number of sexual contacts between the partners since the time of infection of the index case. The estimation of the distribution of the time $T$ from infection of the index case until infection of the case's sexual partner is of interest, with $T$ measured as the number of sexual contacts. Data on the following variables for the partners were available.

serostat: infection indicator of female partner at monitoring (interview) time.

length: monitoring time $C$ measured by the number of sexual contacts since infection of the index case.

age: female partner's age in years at the time of recruitment .

bleeding: indicator of reported bleeding during intercourse prior to recruitment (1=yes).

stdhx: indicator of history of other STDs (1=yes).

nocondom: indicator of any condom use (1=no).

For estimation of $F(c \mid L) = E(\Delta \mid C = c, L)$ we specified a linear logistic regression model with covariates "length", bleeding, condoms, STD, bleeding*condoms, bleeding*STD. A GAM-analysis in Splus showed that deviations from this model were not significant. Secondly, an analysis of the Cox-proportional hazards model for $C$ and the covariates showed that the covariates were independent of $C$. We decided, based on the experience in the simulation experiments, to estimate $g(c \mid x)$ with a kernel density estimator with cross-validated bandwidth.

We estimated the truncated mean $\int_0^\tau (1 - F(t))dt = \int_0^\tau t dF(t) + \bar{F}(\tau)\tau$ with $\tau$ the 0.975-quantile of the $C_i$'s, using the NPMLE $\hat{\mu}_{NPMLE}$ that ignores data on the covariates and the one-step estimator. We found $\tau = 2334.0$, $\hat{\mu}_{NPMLE} = 1602.8$ and $1 - step = 1579.4$. The variance of the influence curve for $\hat{\mu}_{NPMLE}$ and the one-step were estimated as 3827670 and 3551342, respectively. Thus a 0.95-asymptotic-confidence interval for the truncated mean based on the one-step estimator is given by:

$$1579.4 \pm 1.96\sqrt{3551342/88} = 1579.4 \pm 393.8 = [1185.6, 1973.2].$$

When we repeated the analysis with $\tau$ now being the 0.85-quantile of the $C_i$'s we obtained: $\tau = 980$, $\hat{\mu}_{NPMLE} = 788.7$ and $1 - step = 760$. The corresponding variances of the influence curves of $\hat{\mu}_{NPMLE}$ and $1 - step$ were 250794.6 and 215325.8, respectively. A 95% confidence interval based on the one-step estimator is now given by:

$$760 \pm 97 = [663, 857].$$

As expected estimation of the tail created extra variability. Therefore, one obtains more precise estimates of (strongly) truncated means.

## 8 Conclusion and discussion.

The above simulation experiments show that the practical performance of the proposed one step estimator is in keeping with its excellent theoretical properties. Further simulation studies in van der Laan and Hubbard (1997) suggest this performance is maintained even at sample sizes smaller than the ones used in our simulation experiments. Furthermore, van der Laan and Hubbard showed that this approach provides excellent estimates of the distribution function

$F(t)$ by specifying $r(c) = K(\{c - t\}/h)/h$ for some kernel $K$ and bandwidth $h$. This one-step estimator of $F(t)$ uses the information contained in the covariates $L$ in a locally optimal way and, in simulation experiments performed remarkably well. If $C$ is discrete at $t$, then our one-step estimator still yields excellent estimates of $F(t)$ by setting $r(c) = I(c = t)$ and letting $g_n(t \mid L)$ be an estimator of the conditional probability that the sacrificing time equals $t$, given $L$.

The methods presented here can be also used to construct highly efficient estimators for interval censored data with several monitoring times, in the presence of covariates (see van der Laan, Hubbard, 1997). Generalization of our methods to estimate regression parameters with current status data in proportional hazards and linear regression models will be reported elsewhere.

## References


Barlow, R.E. Bartholomew, D.J. Bremner, J.M. and Brunk, H.D. (1972), *Statistical inference under order restrictions*, John Wiley, New York.

Bickel, P.J., Klaassen, A.J., Ritov, Y. and Wellner, J.A. (1993), *Efficient and adaptive inference in semi-parametric models*, Johns Hopkins University Press, Baltimore.

Diamond, I.D. McDonald, J.W. and Shah, I.H. (1986), Proportional hazards models for current status data: application to the study of differentials in age at weaning in Pakistan, *Demography* **23**, 607–620.

Diamond, I.D. and McDonald, J.W. (1991), The analysis of current status data, *Demographic Applications of Event History Analysis*, J. Trussell, R. Hankinson, and J. Tilton (eds.), Oxford: Oxford University Press.

van de Geer, S. (1994), *Asymptotic normality in mixture models,* preprint University of Leiden, the Netherlands.

Gill, R.D., van der Laan, M.J. and Robins, J.M. (1997), Coarsening at Random: Characterizations, Conjectures and Counter-examples, *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*, 255-295, Springer, D.Y. Lin and T.R. Fleming (eds).

Groeneboom, P. and Wellner, J.A. (1992), *Information bounds and nonparametric maximum likelihood estimation*, Birkhäuser Verlag.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Heitjan, D.F. and Rubin, D.B. (1991), Ignorability and coarse data, *Annals of Statististics* **19**, 2244–2253.

Huang, J. (1996), Efficient estimation for the proportional hazards model with interval censoring, *Annals of Statististics*, **24**, 540–568.

Huang, J. and Wellner, J.A. (1995), Asymptotic normality of the NPMLE of linear functionals for interval censored data, case I, *Statistica Neerlandica* **49**, 153–163.

Jacobsen, M. and Keiding, N. (1995) , Coarsening at random in general sample spaces and random censoring in continuous time, *Annals of Statistics*, **23**, 774–786.

Jewell, N.P. Malani, H.M. and Vittinghoff, E. (1994), Nonparametric estimation for a form of doubly censored data with application to two problems in AIDS, *Journal of the American Statistical Association*, **89**, 7–18.

Jewell, N.P. and Shiboski, S.C. (1990), Statistical analysis of HIV infectivity based on partner studies, *Biometrics*, **46**, 1133-1150.

Keiding, N. (1991) Age-specific incidence and prevalence (with discussion), *Journal of the Royal Statistical Society Ser. A*, **154**, 371–412.

van der Laan, M.J. (1997), Locally Efficient Estimation with Current Status Data. *1996 Proceedings of the Biometrics Section of the Annual Meeting of the ASA, Chicago*, 41–49.

van der Laan, M.J., Bickel, P.J. and Jewell, N.P. (1997), Doubly and singly censored current status data: Estimation, regression and asymptotics, *Scandinavian Journal of Statistics*, **24**, 289–308.

van der Laan, M.J. and Robins, J.M. (1997), *Locally efficient estimation with current status data and time-dependent covariates*, Technical report # 64, Division of Biostatistics, University of California, Berkeley.

van der Laan, M.J. and Hubbard, A. (1997), Estimation with Interval Censored Data and Covariates. *Lifetime Data Analysis* **3**, 77–91.

Padian, N., Marquis, L., Francis, D.P., Anderson, R.E., Rutherford, G.W., O'Malley, P.M. and Winkelstein, W. (1987), Male-to-Female transmission of Human Immunodeficiency Virus, *Journal of the American Medical Association*, **258**, 788–790.

Pollard, D. (1990), *Empirical Processes: Theory and applications*, Regional conference series in probability and statistics **2**, Inst. Math. Statist., Hayward, California.

Ritov, Y. and Wellner, J.A. (1988), Censoring, martingales, and the Cox model, *Contemporary mathematics*, **80**, 191–219. American Mathematical Society, Providence.

Robins, J.M. and Rotnitzky, A. (1992), Recovery of information and adjustment for dependent censoring using surrogate markers, *Aids Epidemiology, Methodological issues*, Birkhäuser.

Robins, J.M. (1993), Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers, *Proceedings of the Biopharmaceutical section, American Statistical Association*, 24–33.

Robins, J.M. (1996), Locally efficient median regression with random censoring and surrogate markers, *Lifetime Data: Models in Reliability and Survival Analysis*, Ed. N.P. Jewell et al., Kluwer Academic Publishers, 263–274.

Robins, J. M. and Ritov, Y. (1997), Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models, *Statistics in Medicine*, **16**, 285–319.

Rossini, A.J. and Tsiatis, A.A. (1996) A semiparametric proportional odds regression model for the analysis of current status data, *Journal of the American Statistical Association*, **91**, 713–721.

Shiboski, S.C. and Jewell, N.P. (1992), Statistical Analysis of the time dependence of HIV infectivity based on partner study data, *Journal of the American Statistical Association*, **87**, 360–372.

Sun, J. and Kalbfleisch, D. (1993), The analysis of current status data on point processes, *Journal of the American Statistical Association*, **88**, 1449–1454.

van der Vaart, A.W. and Wellner, J.A. (1996), *Weak convergence and empirical processes*, Springer Verlag.