University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2002 *Paper* 123

Recurrent Events Analysis in the Presence of Time Dependent Covariates and Dependent Censoring

Maja Miloslavsky* Sunduz Keles†
Mark J. van der Laan‡ Steve Butler**

*Division of Biostatistics, School of Public Health, University of California, Berkeley
†Division of Biostatistics, School of Public Health, University of California, Berkeley
‡Division of Biostatistics, School of Public Health, University of California, Berkeley
**Genentech, Inc., South San Francisco, CA

Recurrent Events Analysis in the Presence of Time Dependent Covariates and Dependent Censoring

Maja Miloslavsky, Sunduz Keles, Mark J. van der Laan, and Steve Butler

Abstract

Recurrent events models have lately received a lot of attention in the literature. The majority of approaches discussed show the consistency of parameter estimates under the assumption that censoring is independent of the recurrent events process of interest conditional on the covariates included into the model. We provide an overview of available recurrent events analysis methods, and present an inverse probability of censoring weighted estimator for the regression parameters in the Andersen-Gill model that is commonly used for recurrent event analysis. This estimator remains consistent under informative censoring if the censoring mechanism is estimated consistently, and generally improves on the naive estimator for the Anderson-Gill model in the case of independent censoring. We illustrate the bias of ad hoc estimators in the presence of informative censoring with a simulation study and provide a data analysis of recurrent lung exacerbations in cystic fibrosis patients when some patients are lost to follow up.

1 Introduction

Modeling the occurrence of recurrent events has been a much discussed topic in the last few years. The topic is very important from the medical point of view since many medical outcomes are recurrent. As we show in our application section, our concern with recurrent events arises from the recurrent lung exacerbations in cystic fibrosis patients. Our application also motivated our special concern with large number of recurrent events, time dependent covariates and possibly dependent censoring. In the rest of this section we establish notation and review models commonly used for recurrent events. In the sections that follow, we present estimating functions that account for dependent censoring in the marginal Anderson-Gill multiplicative intensity model, practical issues in applying this approach to recurrent events models, and the application to recurrent lung exacerbations in cystic fibrosis patients. Then we generalize the proposed methodology of accounting for depending censoring to proportional rates model.

Let $(0, \tau]$ be the time period of interest or the study time interval. We refer to τ as the end time and denote full data random variable with $\bar{X}(\tau)$ where $\bar{X}(\tau) = \{X(s) : s \leq \tau\}$ and $X(s)$ is a multivariate process evolving in time. The full data random variable, $\mathbf{X} = \bar{X}(\tau)$, stands for everything that can be observed on a randomly selected subject in the interval $(0, \tau]$ if the subject is not subject to censoring. In particular, we can write $\bar{X}(\tau) = \{\bar{N}(\tau), \bar{Z}(\tau)\}$ where $\bar{N}(t) = \{N(s) : s \leq t\}$ and

$$N(t) = \sum_k I(T_k \leq t)$$

is the recurrent events counting process of interest where T_k stands for the time of k^{th} event. $\bar{Z}(\tau)$ is the set of all the covariate processes collected from the beginning until the end of the study.

In recurrent events analysis, the interest usually lies in modeling the occurrence of recurrent events conditional on covariates so that inference could be drawn about the effect of covariates on the recurrent events process. The main difference between various methods used in literature is the quantity modeled, or the parameter of interest. The model then chosen for the parameter of interest often resembles the Andersen-Gill multiplicative intensity model (Andersen and Gill (1982)). Before we describe our full data model, we will now review some of these parameters of interest and the models used to describe them.

The *intensity* of $N(t)$ is defined as

$$E(dN(t) \mid \bar{X}(t-)) = Y_\lambda(t)\, \lambda(t) \tag{1}$$

where $Y_\lambda(t)$ is an "at risk" indicator defined by the full data random variable $\bar{X}(t-)$ which is the full data up until time $t-$. $\lambda(t)$ is the instantaneous probability of process $N(t)$ jumping at time t conditional on the full data past $\bar{X}(t-)$. Most commonly used model for the intensity of a continuous counting process is the *Andersen-Gill multiplicative intensity model* that is described in great detail in Andersen and Gill (1982); Gill (1984); Andersen et al. (1993) and is given by

$$E(dN(t) \mid \bar{X}(t-)) = Y_\lambda(t)\, \lambda(t) = Y_\lambda(t)\, \lambda_0(t) \exp(\beta\zeta(t)), \tag{2}$$

where $\lambda_0(t)$ is baseline intensity function at time t that is positive and usually left completely unspecified, β is a vector of regression coefficients and $\zeta(t)$ is a known function of the full data past $\bar{X}(t-)$. In the case of recurrent events one will always include the past of the process of interest since the intensity of $N(t)$ at time t will almost always depend on the past of $N(t)$. Surprisingly many authors wrongly reflect on the possibility of using the Andersen-Gill model *without* modeling the dependence of $dN(t)$ on $\bar{N}(t-)$. The conclusion following is that this approach is not acceptable since it assumes independence while Andersen and Gill (1982) do not suggest ignoring $\bar{N}(t)$ when modeling the intensity of the counting process of interest (see also Andersen et al. (1993)). If the intensity is modeled as independent of the past of the process itself, it follows from the definition of the intensity that recurrent events are assumed to be independent. Including the past of $N(t)$ means that we need to specify the dependence among recurrent events in a precise way. Striving to avoid the specification of this dependence structure led to the development of alternate models of recurrent events occurrence that we discuss briefly here and detailed descriptions are given in Wei et al. (1989); Pepe and Cai (1993); Lawless and Nadeau (1995); Lin et al. (2000). Another way to describe the intensity of the process would be to employ frailty models. While we do not consider this option in this work, details for these type of models are given in Andersen et al. (1993) and Oakes (1992).

Wei et al. (1989) propose modeling the marginal hazard of the k^{th} event using a proportional hazards model. Therefore, their parameters of interest are

$$E(dN_k(t) \mid \mathcal{F}^k_{t^-}) \qquad \text{for } k = 1, \ldots, K$$

where $dN_k(t) = I(T_k \leq t)$ is an event specific counting process, $\mathcal{F}^k_{t^-}$ is the event specific history that does not include *any* information on counting processes other than $N_k(t)$, and K is the total number of recurrent events. These marginal intensities allow a subject to be at risk of having k^{th} event without having experienced the $k-1^{\text{st}}$ event, and this makes these approaches hard to interpret. We also note that when the total number of recurrent events K is large, the approach is cumbersome. Finally, drawing inference about the effect of covariates on the true counting process of interest $N(t)$ is not possible. Pepe and Cai (1993) get around the problem of being at risk of having k^{th} event without having experienced $k-1^{\text{st}}$ event by including $\bar{N}_{k-1}(t-)$ in $\mathcal{F}^k_{t^-}$. Their parameter of interest is thus

$$E(dN_k(t) \mid N_{(k-1)}(t-) = 1, \mathcal{F}^k_{t^-}) \qquad \text{for } k = 1, \ldots, K.$$

They also propose to include into $\mathcal{F}^k_{t^-}$ only a subset of covariates that is of interest and to model this quantity using a proportional hazard type of model with event specific baseline hazards.

Lawless and Nadeau (1995); Lawless et al. (1997); Lawless (1995); Lin et al. (2000) all opt to model the rate of recurrent events using the proportional rates model. In their approach, the parameter of interest is the *rate* of $N(t)$ that is defined as

$$E(dN(t) \mid \bar{Z}^*(t-)) = Y_m(t)\, m(t)$$

where $Y_m(t)$ is an "at risk" indicator and $\bar{Z}^*(t)$ is a subset of the full data covariate process $\bar{Z}(t)$. $m(t)$ is the rate of jump occurrence in the process $N(t)$ conditional on some subset

of covariates $\bar{Z}^*(t)$. The important thing to note is that $\bar{Z}^*(t-)$ does not include the counting process history $\bar{N}(t-)$. Lin et al. (2000) prove the asymptotic properties for the *proportional rates model* that is given by

$$E(dN(t) \mid \bar{Z}^*(t-)) = Y_m(t)\, m(t) = Y_m(t)\, m_0(t) \exp(\beta\gamma^*(t)) \tag{3}$$

where $m_0(t)$ is a non-negative baseline rate function that is left unspecified, β is a vector of regression coefficients, and $\gamma^*(t)$ is a known function of $\bar{Z}^*(t-)$. The proportional rates model is sometimes also referred to as proportional means model. If the rate of interest is conditional only on time *independent* covariates $E(dN(t) \mid Z^*)$, then by integration or summing we can obtain $E(N(t) \mid Z^*)$ that is then modeled by a proportional means model. This logic does not hold in the case of time dependent covariates and it is unclear what quantity is obtained by integrating. It is our opinion that modeling the rate using proportional rates model is a reasonable approach than the previously discussed methods in an application with a large number of recurrent events. Also note that by excluding $\bar{N}(t-)$ and including only baseline covariates, proportional rates model would generate nicely interpretable regression coefficients, and this is a strength of this model. In this paper, we will consider proportional rates model as a full data model together with a marginal Anderson-Gill multiplicative intensity model that we describe next.

Let $\bar{W}(t) = \{\bar{N}(t), \bar{Z}^*(t)\}$ where $\bar{Z}^*(t) \subset \bar{Z}(t)$ and hence consists of part of the full data covariate process $\bar{Z}(t)$. As a full data model we are also interested in the following multiplicative intensity model:

$$E(dN(t) \mid \bar{W}(t-)) = Y_\lambda(t)\lambda(t) = Y(t)\lambda_0(t) \exp(\beta\gamma(t)), \tag{4}$$

where $\gamma(t)$ is a function of $\bar{W}(t-)$, and $Y(t)$ and $\lambda_0(t)$ are defined as in the Anderson-Gill multiplicative intensity model given in (2).

In the real world, we often do not observe full data but its censored version. Let C denote the censoring time and let $A(t) = I(C < t)$ denote the censoring process where $C = \infty$ if C is censored by τ. We will represent the *observed data* random variable with $\mathbf{Y} = (\min(\tau, C), \Delta = I(\tau < C), \bar{X}(\tau \wedge C))$. Then, the observed data is simply the collection of n *i.i.d.* random variables $Y_1, \cdots, Y_n$ from the random variable Y. Our goal is to draw inference about the full data parameter of interest β based on observed data. There is a crucial assumption on the censoring process that needs to hold for us to be able to draw inference about full data parameters of interest based on observed data. The distribution of the observed data $\mathbf{Y}$ is indexed by the full data distribution F_X and the conditional distribution $G(\cdot \mid X)$ of the censoring variable C given X. We refer to $G(\cdot \mid X)$ as the censoring mechanism and sometimes simply denote it with G. We denote the conditional hazard of the censoring mechanism $A(t)$ given the full data X with $\lambda_C(. \mid X) = E(dA(t) \mid \bar{A}(t-) = 0, X)$. If the censoring mechanism is allowed to depend on unobserved components of $\mathbf{X}$, then the full data parameter of interest is not identifiable from the distribution of the observed data. Therefore we assume *coarsening at random* (CAR) stating that given the full data, the censoring event defining the observed data depends only on the observed part of the data. For right censored data this means that

$$\text{CAR: } \lambda_C(t \mid \mathbf{X}) = \lambda_C(t \mid \bar{X}(t)) \qquad \text{for } t < \tau.$$

Coarsening at random was originally formulated by Heitjan and Rubin (1991) and further generalized by Jacobsen and Keiding (1995) and Gill et al. (1997). In general, we refer to Robins and Rotnitzky (1992) and Robins (1993) for the introduction and discussion of this CAR definition for the right censored data. The CAR assumption basically says that given the full data $X = x$, the censoring event defining the observed data $Y = y$ depends only on the observed part of x.

The methodology we will pursue requires, aside from the CAR assumption, $\lambda_C(t \mid \bar{X}(\tau)) = \lambda_C(t \mid \bar{X}(t))$ and the full data marginal multiplicative intensity model assumed, a model for the censoring mechanism. In particular, we will assume a Anderson-Gill multiplicative intensity model for $\lambda_C(t \mid \bar{X}(t-))$ given by

$$\lambda_C(t \mid \bar{X}(t-)) = Y_C(t)\lambda_{0,C}(t)\exp(\beta_C\zeta_C(t)), \tag{5}$$

where $Y_C(t)$ is the "at risk indicator" for censoring, $\lambda_{0,C}(t)$ is unspecified baseline hazard and $\zeta_C(t)$ is a known function of $\bar{X}(t-)$. Moreover, we need an identifiability condition that there exists a $\tau^* \leq \tau$ such that

$$\bar{G}(\tau^* \mid X) = P(C \geq \tau^* \mid X) > 0, F_X - a.e. \tag{6}$$

In the recurrent events data literature, the general approach of dealing with the observed data of recurrent events is through the modeling of the observed data counting process. Since we now are working with the observed instead of the full data, the recurrent event counting process we observe is not the counting process of interest but

$$N^*(t) = \sum_k I(T_k \leq t \wedge C) = N(t \wedge C).$$

Based on the observed data, the intensity we can model is

$$E(dN^*(t) \mid \bar{X}(t - \wedge C), \bar{A}(t^-)) = Y_{\lambda^*}(t)\ \lambda^*(t), \tag{7}$$

where Y_{λ^*} is the risk indicator and $\lambda^*(t)$ is the instantaneous probability of process $N^*(t)$ jumping at time t conditional on the observed past $(\bar{X}(t^- \wedge C), \bar{A}(t))$. We can model $\lambda^*(t)$ once again using the multiplicative intensity model.

Similarly the rate of $N^*(t)$ conditional on $(\bar{Z}^*(t \wedge C), \bar{A}(t))$ can be modeled with the proportional rates model as

$$E(dN^*(t) \mid \bar{Z}^*(t \wedge C), \bar{A}(t-)) = Y_{m^*}(t)m^*(t). \tag{8}$$

The question of interest is now: when are the parameters of the observed data distribution equal to the full data parameters that are of interest? More explicitly, when do we have $\lambda^*(t)$ of observed data counting process equal to $\lambda(t)$ of the full data counting process in model (1). Similarly, when do we have $m^*(t) = m(t)$ in the proportional rates model (8)?

In the marginal Anderson-Gill multiplicative intensity model where the conditioning set is the whole past $\bar{X}(t)$, if CAR holds, then $\lambda^*(t) = \lambda(t)$ and therefore, the intensity of the

observed process is equivalent to the intensity of the full data process that is of interest. The reason for this is the factorization of the density of $\mathbf{Y}$ in a F_X part and $G(\cdot|X)$ part as a result of CAR. The main result emerging from this is that one can estimate the intensity of the observed data process and obtain the full data parameter of interest (Andersen et al. (1993)). In the case of rates, if CAR holds such that $\lambda_C(t \mid \bar{X}(\tau)) = \lambda_C(t \mid \bar{X}(t))$, and moreover if $E[dN(t) \mid \bar{Z}^*(t), C \geq t] = E[dN(t) \mid \bar{Z}^*(t)]$ (equivalent to $\lambda_C(t \mid \bar{X}(t)) = \lambda_C(t \mid \bar{Z}^*(t))$), then $m^*(t) = m(t)$). Note in particular that the second assumption implies that the censoring mechanism is independent of the counting process of interest given the covariates $\bar{Z}^*(t)$. Under these conditions, the parameter of interest can be estimated consistently using observed data partial log-likelihood in the proportional rates model as proposed by Lin et al. (2000). For the marginal multiplicative intensity model given in (4), if $\lambda_C(t \mid \bar{X}(t)) = \lambda_C(t \mid \bar{W}(t))$, then the intensity of the full data counting process can be obtained with a similar approach. However, these independence assumptions can easily be violated in real life situations in the sense that censoring might depend on covariates in $\bar{Z}(t)$ beyond $\bar{Z}^*(t)$ which will violate the independence assumption for the proportional rates model. Similarly, it might depend on covariates beyond $\bar{W}(t)$ and violate this assumption for the marginal multiplicative intensity model. In that case, though the estimators obtained using the observed data partial likelihood are consistent for the observed data model parameters, these parameters differ from the full data parameters of interest. For this reason, we are not following the route of modeling the observed data counting process but directly modeling the intensity of the full data counting process.

The situation where the censoring mechanism depends on covariates that are not in the conditioning set is often referred to as dependent (informative) censoring. In the case of informative censoring, the ad hoc estimation procedures from the observed data will result in inconsistent estimators. The aim of this paper is to propose methods for consistent estimation of the regression parameters in the full data models (4) and (3) from the observed data in the presence of dependent censoring.

We firstly propose a class of observed data estimating functions for the regression parameter β in the marginal marginal Anderson-Gill multiplicative intensity model given in (4). The proposed class of estimating functions are obtained as *inverse probability of censoring weighted (IPCW)* mappings of the full data estimating functions and they remain unbiased in the case of dependent censoring if censoring mechanism is estimated consistently and the identifiability condition (14) holds. We then specify a particular estimating function from this class that reduces to the ad hoc estimating function obtained from the observed data partial likelihood and is typically used for estimating the regression parameters in the Anderson-Gill multiplicative intensity model under independent censoring. This estimating function coincide with the estimating function proposed by Robins (1993) for cox-proportional hazards model which is a special case of marginal Anderson-Gill multiplicative intensity model. The strengths of the proposed estimating function is demonstrated with a simulation study and it is used in a real data example. We then show how the proposed method applies to the proportional rates model. When the censoring is independent of the counting process of the interest conditional on the covariates that are included in the model, our method does not require any different assumptions than Lin et al. (2000)'s method. In other words, the correctness of the estimated censoring mecha-

nism and the identifiability condition gains importance only when censoring is dependent. In addition, for the interested reader, we review the general methodology of doubly robust estimation for censored data problems in the Appendix and provide a doubly robust estimator for our parameter of interest in recurrent events data analysis. This estimator improves on the proposed IPCW estimator and has the potential of staying consistent even when the censoring mechanism is not estimated consistently and the identifiability assumption (14) is violated.

2 Methods

In this section, we will address the estimation of the regression parameters in the full data model (4) based on the observed data. We will firstly review the estimation problem based on the full data. Efficient estimation based on the full data $\bar{X}(t)$ in the marginal Andersen-Gill multiplicative intensity model given by (4) is a solved problem. The general class of full data estimating functions will be provided in the following subsections (from van der Laan and Robins (2002), Lemma 2.2, p.107) and the full data efficient estimating function will be denoted with $S^F_{eff}(. \mid \beta)$. These estimating functions are based on the full data partial likelihood for the marginal Andersen-Gill model and the desirable asymptotic properties of the resulting parameter estimates are obtained using the martingale properties of the estimating functions (Andersen et al. (1993)). We obtain a class of observed data estimating functions from full data estimating functions using IPCW mapping. After deriving this general class, we point out to a particular choice of estimating function that reduces to the ad hoc estimating function obtained from the observed data partial likelihood which has been used when censoring is independent (Lin et al. (2000)).

2.1 Observed data estimating functions for marginal the Andersen-Gill recurrent events full data intensity model

Recall from Section 1 that, in the recurrent events setting we write the full data as $\bar{X}(\tau) = (N(\tau), \bar{Z}(\tau))$ where

$$N(t) = \sum_k I(T_k \leq t)$$

is our recurrent events counting process of interest and $\bar{Z}(\tau)$ is a collection of all the covariate processes. Given that C is the censoring variable, the observed data is $\mathbf{Y} = (\min(\tau, C), \Delta = I(\tau < C), \bar{X}(\tau \wedge C))$.

As we discuss in the introduction, we are interested in modeling the intensity of the full data counting process. Andersen-Gill multiplicative intensity model assumes that

$$E(dN(t)|\bar{W}(t-)) = Y(t)\lambda_0(t)\exp(\beta\gamma(t)),$$

where $\gamma(t)$ is a known function of $\bar{W}(t-)$. While our parameter of interest is the full data counting process of interest, we have observed data available and want to draw inference about the full data parameter based on the observed data. We know that under CAR the

intensity of the observed data process reduces to the intensity of the full data counting process if the conditioning set of the full data intensity model includes the whole past $\bar{X}(t)$. However, since we are not conditioning on $\bar{X}(t)$ but only some subset $\bar{W}(t)$, we need to derive estimating equations for the parameter of interest in this general model. We firstly look at the estimation problem based on the full data.

The class of all full data estimating functions in model (2) is given by (Lemma 2.2 of van der Laan and Robins (2002))

$$\left\{ D_h(. \mid \mu, \lambda_0) = \int \left[h(t, \bar{W}(t-)) - g(h)(t) \right] dM_{\beta,\lambda_0}(t) : h \right\} \tag{9}$$

where $g(h)(t)$ equals

$$g(h) = \frac{E[h(t, \bar{W}(t-))Y(t)\exp(\beta\gamma(t))]}{E[Y(t)\exp(\beta\gamma(t))]},$$

and $dM_{\beta,\lambda_0}(t) = dN(t) - E(dN(t) \mid \bar{W}(t-)) = dN(t) - Y(t)\lambda_0(t)\exp(\beta\gamma(t))$.

The full data partial log-likelihood for the Andersen-Gill model and only one observation can be written as

$$\log \mathcal{L} = \int_0^\tau \log(Y(t)\lambda_0(t)\exp(\beta\gamma(t)))dN(t) - \int_0^\tau Y(t)\lambda_0(t)\exp(\beta\gamma(t))dt.$$

The score for β is given by

$$S_\beta = \frac{\partial}{\partial\beta}\log\mathcal{L} = \int_0^\tau \gamma(t) dM_{\beta,\lambda_0}(t).$$

Moreover, the efficient score is given by (Ritov and Wellner (1988); van der Laan and Robins (2002), Lemma 2.2, p.108)

$$S^F_{eff}(. \mid \beta) = \int_0^\tau \left[\gamma(t) - \frac{E[\gamma(t)Y(t)\exp(\beta\gamma(t))]}{E[Y(t)\exp(\beta\gamma(t))]} \right] dM_{\beta,\lambda_0}(t). \tag{10}$$

$$\tag{11}$$

Note that $S^F_{eff}(. \mid \beta)$ is an element of the class of full data estimating functions given in (9) with $h(t, \bar{W}(t-)) = \gamma(t)$.

Provided that we do not always observe full data $\bar{X}(\tau)$ but its censored version Y we are still interested in finding practical and well behaved estimators of full data intensity model parameters. If the counting process of interest is independent of censoring time C conditional on $\bar{W}(t)$, then the estimating equation given by Andersen et al. (1993) equals

$$S^*_\beta = \int_0^\tau \left[\gamma(t) - \frac{E[\gamma(t)I(t<C)\exp(\beta\gamma(t))]}{E[I(t<C)\exp(\beta\gamma(t))]} \right] I(t<C) dM_{\beta,\lambda_0}(t). \tag{12}$$

This corresponds with the score of the partial likelihood for β and λ_0 ignoring the covariate process beyond $\bar{W}(t)$ and it yields consistent and asymptotically normal estimators. If, however, not all covariates that are relevant to the censoring mechanism are included into the model, this estimating function is biased, hence does not yield consistent estimators.

Thus, we need to map the full data estimating functions into the observed data ones so that the resulting estimators are consistent under a more general censoring model.

A general way of obtaining such consistent estimating functions is to map full data estimating functions into observed data estimating functions using IPCW mapping (Robins and Rotnitzky (1992)).

Let $\Delta(t) = I(C > t)$. Then, a choice of IPCW estimating function is given by

$$U_G(Y \mid D_h) = \int_0^\tau \underbrace{[h(t, \bar{W}(t-)) - g(h)(t)]}_{h^\star(t,\bar{W}(t-))} \frac{dM_{\beta,\lambda_0}(t)\Delta(t)}{\bar{G}(t \mid X)}, \tag{13}$$

where $\bar{G}(t \mid X)$ is $P(C > t \mid X)$. Note that $U_G(. \mid D_h)$ satisfies $E(U_G(Y \mid D_h) \mid X) = D_h(X \mid \mu, \lambda_0)$ under the assumption that

$$P(C > \tau \mid X) > \delta > 0, \tag{14}$$

and hence it yields consistent estimators in the presence of dependent censoring. Note also that this identifiability condition can be arranged by making the integral in the expression of $U_G(. \mid D_h)$ go up to a τ^* such that $P(C > \tau^* \mid X) > \delta > 0, F_X - a.e.$ In this case, the efficiency of the resulting estimator will depend on how close τ^* is to τ since this modification allows the data up to τ^* to be used.

2.1.1 A particular choice of observed data estimating function

We note that for each $h(.)$ one can construct an IPCW type estimating function as in (13). Provided that we model the intensity of interest conditional on $\bar{W}(t-)$, we want to insure that if $\lambda_C(t|\bar{X}(t-)) = \lambda_C(t|\bar{W}(t-))$, then our estimating equation reduces to the naive estimating function given in (12). Practically, this means that we want to ensure that the weighted estimating equations perform at least as well as the "naive" approach.

While it is often convenient to choose $D_h(X|\mu, \lambda_0) = S^F_{eff}(. \mid \beta)$, and we imply this choice in the previous discussion, the following full data estimating function is a more parsimonious choice in the presence of censoring. Define

$$D^*_h(X|\mu, \lambda_0) = \underbrace{\int_0^\tau \left[\gamma(t) - \frac{E[\gamma(t)\bar{G}(t|\bar{W}(t-))Y(t)\exp(\beta\gamma(t))]}{E[\bar{G}(t|\bar{W}(t-))Y(t)\exp(\beta\gamma(t))]}\right] \bar{G}(t|\bar{W}(t-))\, dM_{\beta,\lambda_0}(t)}_{h^*(t,\bar{W}(t-))}.$$

It can be easily verified that $D^*_h(X|\mu, \lambda_0)$ is an element of the class of full data estimating functions given in (9). Applying the time dependent weighting to this full data estimating equation yields the following observed data estimating equation:

$$U_G(Y \mid D^*_h) = \int_0^\tau \left[\gamma(t) - \frac{E[\gamma(t)\bar{G}(t|\bar{W}(t-))Y(t)\exp(\beta\gamma(t))]}{E[\bar{G}(t|\bar{W}(t-))Y(t)\exp(\beta\gamma(t))]}\right] \frac{\bar{G}(t|\bar{W}(t-))\Delta(t)dM_{\beta,\lambda_0}(t)}{\bar{G}(t|X)}, \tag{15}$$

which can be rewritten as

$$U_G(Y \mid D_h^*) = \int_0^\tau \left[\gamma(t) - \frac{E[\frac{I(C>t)}{\bar{G}(t|X)}\gamma(t)\bar{G}(t|\bar{W}(t-))Y(t)\exp(\beta\gamma(t))]}{E[\frac{I(C>t)}{\bar{G}(t|X)}\bar{G}(t|\bar{W}(t-))Y(t)\exp(\beta\gamma(t))]} \right] \frac{\bar{G}(t|\bar{W}(t-))\Delta(t)dM_{\beta,\lambda_0}(t)}{\bar{G}(t|X)}$$

It is straight forward to see that if $\lambda_C(t|X) = \lambda_C(t|\bar{W}(t-))$, then $\bar{G}(t|\bar{W}(t-))/\bar{G}(t|X) = 1$ and $U_G(Y \mid D_h^*)$ reduces to the estimating function given by (12).

Estimating functions weighted in this fashion yield the following expression for the baseline hazard

$$\begin{aligned} \lambda_0(t) &= \frac{E\left[\frac{\Delta(t)\bar{G}(t|\bar{W}(t-))}{\bar{G}(t|X)}dN(t)\right]}{E\left[\frac{\Delta(t)\bar{G}(t|\bar{W}(t-))}{\bar{G}(t|X)}Y(t)\exp(\beta\gamma(t))\right]} \\ &= \frac{E\left[dN(t)\bar{G}(t|\bar{W}(t-))\right]}{E\left[\bar{G}(t|\bar{W}(t-))Y(t)\exp(\beta\gamma(t))\right]}, \end{aligned}$$

which we can obtain by double expectation and conditioning in both expectations on X. Here, we use that

$$\begin{aligned} E(dN(t)\bar{G}(t|\bar{W}(t-))) &= E[\, E\,(dN(t)|\bar{W}(t-))\bar{G}(t|\bar{W}(t-))] \\ &= \lambda_0(t)E(Y(t)\exp(\beta\gamma(t))\bar{G}(t|\bar{W}(t-))). \end{aligned}$$

This suggests the following estimator of λ_0 given an estimator $\hat{G}$ of G

$$\hat{\lambda}_0(t \mid \beta) = \frac{\sum_{i=1}^n \left[\frac{\Delta_i(t)\hat{\bar{G}}(t|\bar{W}_i(t))}{\hat{\bar{G}}(t|X_i)}dN_i(t)\right]}{\sum_{i=1}^n \left[\frac{\Delta_i(t)\hat{\bar{G}}(t|\bar{W}_i(t))}{\hat{\bar{G}}(t|X_i)}Y_i(t)\exp\left(\beta\gamma_i(t)\right)\right]} \tag{16}$$

Given estimators $\widehat{h}^*, \widehat{G}, \widehat{\lambda}_0$ of h^*, G, λ_0, we can obtain an estimator for β by solving the following estimating equation

$$0 = \sum_{i=1}^n U_G(Y_i \mid \widehat{G}, \hat{D}_h^*(. \mid \beta, \widehat{\lambda}_0)).$$

One can estimate G by fitting a multiplicative intensity model given in (5) for the censoring process. Then h^* can simply be estimated by substituting $\widehat{G}$ for G and estimating the expectations empirically. In summary, the proposed estimating function remains consistent and asymptotically normal under dependent censoring if $\widehat{G}$ is a consistent estimator of G and G satisfies the identifiability assumption (14), and it reduces to the naive estimating equation if censoring is independent. Note in particular that if censoring is independent of the covariates we are conditioning on we will have $\bar{G}(. \mid \bar{W}(t))/\bar{G}(. \mid X) \simeq 1$, and hence the estimator obtained will not be affected by the correctness of the assumed model for the censoring mechanism and the identifiability condition (14). Moreover, if

one always estimates the weights even in the case of independent censoring, the resulting estimator is more efficient than the naive estimator (van der Laan and Robins (2002), Theorem 2.3, p.135).

One of the strengths of this weighted estimating equation is that it can easily be implemented by using `coxph()` routine of S-plus. This routine for fitting Andersen-Gill multiplicative intensity model accepts weights of the form $\Delta(t)w(t)$. In our application we set $w(t) = \widehat{\bar{G}}(t \mid \bar{W}(t-))/\widehat{\bar{G}}(t \mid X)$.

The standard errors one obtains from the `coxph` software will be conservative since `coxph` treats the weights as known where as in truth we are estimating the weights by substituting $\widehat{\bar{G}}(.)$. However, one can still use these standard errors to get conservative confidence intervals of the regression parameters. In order to obtain the correct confidence intervals, one needs to estimate the correct standard errors either by bootstrap or using the influence curve approach of van der Laan and Robins (2002) (Lemma 3.2, p.192).

3 Simulation study

We have done a simulation study to assess the finite sample performance of our inverse weighted estimator. In all of the simulations, the number of observations are set to $N = 200$.

3.1 With Time Independent Covariates

Consider a study where each subject is randomly assigned a treatment arm of interest. We will denote the treatment variable by A and $A \in \{0, 1\}$. Suppose that the goal of this study is to estimate the causal effect of treatment A on the survival time T. Let $N(t) = I(T \leq t)$ and $X(t) = (N(t), A, Z)$ where Z is a baseline covariate. Since we are interested the effect of treatment A on the survival time we have $\bar{W}(t) = (N(t), A)$ and

$$E(dN(t) \mid \bar{W}(t-)) = I(T \geq t)\lambda(t \mid A)dt,$$

where $\lambda(t \mid A)dt = P(T \in dt \mid T \geq t, A)$ is the hazard of failure within the treatment group. Assuming a multiplicative intensity model, which is the cox-proportional hazards model in this special case, we have

$$\lambda(t \mid A) = \lambda_0(t) \exp(\beta_0^* + \beta_1^* A). \tag{17}$$

The causal parameter of interest is the regression coefficient β_1^* in front of A. An ad hoc method for estimation of β would be to fit a cox-proportional hazards model for the right censored data on T ignoring covariates beyond A. However, if C is not independent of T given A, then this estimator is inconsistent. It is not hard to imagine possible scenarios when censoring depends on covariates beyond A. For example, in this clinical trial, people might drop out of the treatment because of possible side effects of the treatment on subjects

with certain Z measurements. In addition, this ad hoc method will be very inefficient, even when C is known to be independent of T, given A (Robins (1993)).

To mimic such a study we generated data as follows:

- Generate the treatment $A \sim Bernoulli(p)$ and the baseline covariate $Z \sim \mathcal{N}(0,1)$.
- Generate T from $\lambda_T(t) = \lambda_{0,T}(t) \exp(\beta_1^t A + \beta_2^t Z)$, where $\lambda_{0,T}(t)$ is the hazard from the truncated exponential distribution with parameters λ_t and τ.
- Generate C from $\lambda_C(t) = \lambda_{0,C}(t) \exp(\beta_1^c A + \beta_2^c Z)$, where $\lambda_{0,C}(t)$ is the hazard from the exponential distribution with parameter λ_c.

The hazard of truncated exponential distribution is used for baseline hazard of the event times, and the hazard of the exponential distribution is used for the censoring baseline hazard so that $\bar{G}(T \mid X) > \delta > 0, F_X - a.e.$ is satisfied.

The observed data obtained from this simulation is $Y_i = ((T_i \wedge C_i), \Delta_i = I(T_i \leq C_i), A_i, Z_i), i = 1, ..., N$. We are interested in the effect of treatment A on the hazard of survival. So the parameter of interest is the regression coefficient β_1^* in the model (17). Note that $\beta_1^t \neq \beta_1^*$, thus model (17) is misspecified which is common in real data applications. We then define the parameter of interest as β_1^* for which one obtains the best approximation to the true hazard of survival time T conditional on treatment A using model (17). One finds this by setting β_1^* equal to the maximum likelihood estimator of it in model (17) based on a large number of observations. Hence, we obtain a good estimate of the true parameter β_1^* by generating a large number of observations (e.g. $N = 1000000$) (T, A, Z) from the data generating distribution and fitting the model (17) with `coxph` using the full data. Estimate of β^* obtained in this method corresponds to the minimizer of the Kullback Leibner projection of the true data generating distribution on the model (17).

Results of this simulation study are summarized in Table 1. We see from this table that ignoring the dependence of the censoring on covariates other than the covariate of interest in the model causes serious bias even with low censoring percentages. The results become dramatically bad when censoring percentage increases.

3.2 With Time Dependent Covariates

In this simulation study, we generated event times from a logistic distribution with discrete support based on a baseline covariate A that represents the treatment assignment and a time dependent covariate Z. We summarize the data generation process as follows:

- Generate the treatment variable $A \sim Bernoulli(p)$ and the baseline covariate $Z \sim Gamma(1,1)$. This value of the Z corresponds with the value of the time dependent covariate at $t = 0$.

10% Censoring		
	Unweighted	Weighted by $\frac{\Delta(t)\bar{G}(t\|A)}{\bar{G}(t\|A,Z)}$
$\widehat{Bias}$	0.2216611	0.0262708
$\widehat{MSE}$	0.0794047	0.0279184
25% Censoring		
	Unweighted	Weighted by $\frac{\Delta(t)\bar{G}(t\|A)}{\bar{G}(t\|A,Z)}$
$\widehat{Bias}$	0.4379876	0.093323
$\widehat{MSE}$	0.2306377	0.046926
50% Censoring		
	Unweighted	Weighted by $\frac{\Delta(t)\bar{G}(t\|A)}{\bar{G}(t\|A,Z)}$
$\widehat{Bias}$	0.670974	0.0014334
$\widehat{MSE}$	0.5097888	0.2342156

Table 1: *With time independent covariates:* Simulation results on bias and mean squared errors of the two estimators for the regression parameter β_1^* based on 2000 replicates. Samples of size 200 are generated with right censoring percentage 10%, 25%, and 50%. β_1^* equals 0.616 based on $N = 1000000$ observations. The parameters of the data generating distributions are set as follows: $\beta_1^t = 4$, $\beta_2^t = 5$, $\tau = 10$, $\lambda_t = 0.01$, $\beta_1^c = 1$, $\beta_2^c = 5$. λ_c is set to 0.06, 0.2, and 1.2 for censoring proportions 10, 25 and 50, respectively.

- Generate T : Starting from $t = 0+$, perform the following two steps at each $t \in \{1, \cdots, 52\}$
  1. Compute the value of the time dependent covariate $Z(t) = Z * t$
  2. Draw a 0-1 variable from the following logistic distribution

$$P(T = t \mid T >= t, A, Z(t)) = logit(\beta_0^t + \beta_1^t A + \beta_2^t Z(t))$$

  until a 1 is drawn at a t_i. Set $T = t_i$.

- Generate C : Starting from $t = 0+$, perform the following step at each $t \in \{1, \cdots, 52\}$
  1. Draw a 0-1 variable from the following logistic distribution

$$P(C = t \mid C >= t, A, Z(t)) = logit(\beta_0^c + \beta_1^c A + \beta_2^c Z(t))$$

  until a 1 is drawn at a t_j. Set $C = t_j$.

As in the time independent simulation, the observed data is $Y_i = ((T_i \wedge C_i), \Delta_i = I(T_i \leq C_i), A_i, \bar{Z}_i(T_i \wedge C_i)), i = 1, ..., N$ and we are interested in the effect of treatment A on the hazard of survival. The parameter of interest is the regression parameter in the model

$$E(dN(t) \mid A) = I(T \geq t)\lambda(t \mid A)dt = I(T \geq t)\lambda_0(t)\exp(\beta_0^* + \beta_1^* A),$$

and we set its true value to the full data maximum likelihood estimator based on a large number of uncensored observations (as in the time independent study). The results of these time dependent simulations are summarized in Table 2. We see that the weighted

10% Censoring		
	Unweighted	Weighted by $\frac{\Delta(t)\bar{G}(t\|A)}{\bar{G}(t\|A,Z)}$
$\widehat{Bias}$	0.02937382	0.01048841
$\widehat{MSE}$	0.03262307	0.03168958
20% Censoring		
	Unweighted	Weighted by $\frac{\Delta(t)\bar{G}(t\|A)}{\bar{G}(t\|A,Z)}$
$\widehat{Bias}$	0.04165596	0.01750265
$\widehat{MSE}$	0.04558592	0.03198636

Table 2: *With time dependent covariates:* Simulation results on bias and mean squared errors of the two estimators for the regression parameter β_1^* based on 1000 replicates. Samples of size 200 are generated with right censoring percentage 10%, 20%. β_1^* equals 1.35 based on $N = 100000$ observations. The parameters of the event time generating distribution are set as follows: $\beta_1^t = -4$, $\beta_1^t = 2$, $\beta_2^t = 2$. Parameters of the censoring time generating distribution are $\beta_0^c = -1.4$, $\beta_1^c = -3.2$, $\beta_2^c = 3$ for 10% censoring; $\beta_0^c = 0.7$, $\beta_1^c = -3.2$, $\beta_2^c = 3$ for 20% censoring.

estimator outperforms the naive unweighted estimator at both of the censoring proportions (10%, 20%). However, the difference in bias is not as dramatic as it was in the time independent covariate scenario probably because using $Z(t) = Zt$ caused less informative censoring than using $Z(t) = Z$.

4 Recurrent exacerbations in cystic fibrosis patients

Cystic fibrosis (CF) is the most common genetic disease in the US. The disease is the result of a mutation in a membrane protein that functions as a chloride ion channel and is therefore indirectly responsible for water movement across the cell membrane. The main effect of this mutation is thick and viscous mucus that is produced by cells retaining water. This mucus leads to complications in epithelial tissues causing digestive problems and most importantly, lung disease leading to respiratory failure, the most common cause of death in CF patients.

The Epidemiologic Study of Cystic Fibrosis (ESCF) described in Morgan et al. (1999) is a multi-center observational study prospectively collecting information on cystic fibrosis patients involved. ESCF also serves as a phase-IV observational study of dornase alfa use (Pulmozyme, Genentech Inc., South San Francisco, CA), and it is funded by Genentech, Inc. The study is ongoing and enrollment started in December 1993. There have been over 20,000 patients enrolled. The data are collected at all clinic visits and hospitalizations,

and consist of demographic information, medical conditions, lung function, microorganism presence, routine and antibiotic therapies, adverse events and discontinuation data.

The progression of CF is best described by the decline in lung function and the occurrence of lung infections or exacerbations. Most severe lung exacerbations are treated with IV antibiotics in a hospital. CF patients experience four exacerbations per year on average and as patient's health deteriorates, exacerbations get more frequent. We are interested in modeling the occurrence of lung exacerbations since their frequency is one of the main indicators of disease's progression. Since the occurrence of exacerbations depends on various factors and stages of the disease, we expect our intensity regression models to describe the relationship of exacerbation occurrence with: lung function as measured by spirometries, presence of specific microorganisms in the lung mucus, and other clinically important covariates.

In the analysis we present here, we focus on the occurrence of IV treated exacerbations in prepubescent patients. This means that we are only concerned with the occurrence of IV treated exacerbations in patients between 6 and 14 years of age. Since we know that the frequency of exacerbations increases with increasing age and are not interested in estimating the effect of age, we set up our analysis so that age is our time scale. We set up a series of inclusion/exclusion criteria to more closely define our population of interest. The entry into the study is defined as the age at which a patient already has had two respiratory cultures examined and has had a measurement of pulmonary function done while healthy. Patients can "enter" the analysis between ages 6 and 12, and have to have all the necessary baseline information available at entry. Since we are interested in prepubescent patients, we discontinue them from our analysis once they reach age 14.

Based on these inclusion/exclusion criteria, the extracted ESCF young patients data set yields 4855 qualifying patients, 51% of which are female. Among these patients, average follow up is almost 3 years with maximum being 5.75 and minimum one month. 38.7% of patients enter the study between ages 6 and 7, 13.3% between ages 7 and 8, and the rest are evenly distributed among the remaining age groups. In our intensity model, we consider the following covariates as possibly relevant subset of the observed past. FEV1, forced expiratory volume in one second that is expressed as a percent of predicted for given sex, age and height, is measured via spirometry at every patient visit and is our main measure of lung function. Since presence of some microorganisms in the respiratory tract indicates the stage of the disease, we consider time dependent indicator variables showing if the microorganism was present in the last respiratory culture done or not. The microorganisms we consider are: Burkoholderia Cepacia or B. Cepacia, Pseudomonas Aeruginosa or P.A., Stenotrophomonas Maltophilia or X. Maltophilia and Candida. We also consider covariate P.A. EVER in our models. This covariate indicates if a given patients has ever had a respiratory culture positive for P.A. and is important since it is still unclear if P.A. once it appears can be eradicated from the respiratory tract or if it marks a new stage of the disease. Other important covariates include growth and development status of a patient as well as other health status indicators. Since CF patients have problems absorbing nutrients, their growth is somewhat slower than in healthy children. To capture their status we consider as covariates weight for age percentile (WTPCTA)

and height for age percentile (HTPCT). We also consider the level of sputum productivity (SPUTMACT) and cough frequency (COUGHFRQ) as indicators of severity of the disease at a given time both of which are categorical variables with three levels. We also consider SEX of a patient as a covariate and attempt to model the dependency between past and current exacerbations. To model the dependency, we consider as covariates $N(t^-)$ or TOT, indicating the total number of exacerbations before time t, ALAST, the age at last exacerbation, and BEGAGE indicating age at the beginning of study or observation that should adjust for the seemingly unfortunate use of TOT as covariate since TOT corresponds not only to true number of observations that occurred since age 6 but since the age of entry into the analysis data set. It is important to note that all the covariates with the exception of SEX are time dependent and are usually measured at every clinic visit. This set of covariates is the subset of the full history that we refer to as $W(t)$ in the previous section.

We define the full data for subject i as everything that can be observed on the subject between the age of entry into the study C_{li}, and the age at the end of the study C_{Ei}. We denote full data as $X(C_{li}, C_{Ei})$. The age at the end of study is defined as the minimum of age on December 1, 1999, age 14, age of death, and age of lung transplant. Since patients die due to progression of their illness and lung transplant is usually granted once CF has advanced to the stage of endangering the life of the patient, the two events, death and transplant, can be considered statistically equivalent. It is important to note that death and/or transplant are not censoring events since the process describing recurrent lung exacerbations can no longer jump and is no longer of interest once death occurs. Therefore, death or transplant define the end point for full data as does the end of the study. Provided this definition of full data we define the counting process of interest as before

$$N_i(t) = \sum_k I(T_{ki} \leq t)$$

where T_{ki} is the time of k^{th} event for individual i.

8% of the individuals in the data set are censored due to loss to follow up. We denote the age of right censoring for individual i with C_{ri}. Then the observed data for individual i can be represented as $Y_i = X(C_{li}, C_{Ei} \wedge C_{ri})$ and the observed counting process is then

$$N_i^*(t) = \sum_k I(T_{ki} \leq t \wedge C_{ri})$$

As we discuss previously, when we restrict conditioning on $\bar{X}(t)$ to the subset $\bar{W}(t)$, in order to assure the consistency of our full data parameter estimates based on observed data, we need to have in addition to CAR that $\lambda_C(t \mid \bar{X}(t)) = \lambda_C(t \mid \bar{W}(t))$, which implies that the censoring is independent of the counting process of interest given the covariates we are conditioning on. If we assume that this condition is met, we fit Andersen-Gill model without weights. After going through a model selection procedure including selection of covariates, checking of the functional form and examining of residuals, we obtain estimated coefficients given in Table 3. TOT1, TOT2, ..., TOT8 correspond to indicators for events. The indicator variables created correspond to events TOT=1, TOT=2,...,TOT $\geq$ 8. Now

Variable	Estimate	std.err.	Wald z	P-value
SEX	0.0888	0.0252	3.52	0.0004
BEGAGE	0.4056	0.0147	27.55	0.0000
FEV1	-0.0105	0.0006	-17.71	0.0000
TOT1	0.8441	0.0370	22.81	0.0000
TOT2	1.3420	0.0441	30.40	0.0000
TOT3	1.6388	0.0512	31.99	0.0000
TOT4	1.8515	0.0594	31.18	0.0000
TOT5	2.1924	0.0660	33.20	0.0000
TOT6	2.1697	0.0753	28.82	0.0000
TOT7	2.4739	0.0826	29.94	0.0000
TOT8	2.5547	0.0609	41.92	0.0000
P.A.	0.2340	0.0349	6.69	0.0000
P.A. EVER	0.2738	0.0486	5.63	0.0000
B. CEPACIA	0.3730	0.0661	5.65	0.0000

Table 3: Estimated coefficients for the Andersen-Gill model of exacerbation intensity assuming independent censoring.

we consider the possibility that right censoring depends is not independent of the process of interest conditional on the covariates included into the model so that only CAR holds but that it is not necessarily true that $\lambda_C(t \mid \bar{X}(t)) = \lambda_C(t \mid \bar{W}(t))$. If only CAR holds without the additional assumption, then our previously obtained estimates of full data parameters based on observed data are biased. However, the strength of the association of censoring with covariates, the prevalence of censoring, and the covariates included into the model of interest, all affect how different the estimated coefficients obtained by ignoring dependent censoring are in practice from the coefficients obtained using the described estimating function approach. Since we can not judge the extent of dependent censoring effect, it is advisable to at least use estimating equations that are unbiased in the presence of censoring.

In the analysis of ESCF recurrent lung exacerbations data in the presence of censoring, we use the time dependently weighted estimating equation $U_G(. \mid D_h^*)$. This estimating function is given by

$$U_G(. \mid D_h^*) = \int \left[\gamma(t) - \frac{E[\frac{I(C>t)}{\bar{G}(t|X)}\gamma(t)\bar{G}(t|\bar{W}(t-))Y(t)\exp(\beta\gamma(t))]}{E[\frac{I(C>t)}{\bar{G}(t|X)}\bar{G}(t|\bar{W}(t-))Y(t)\exp(\beta\gamma(t))]} \right] \frac{\bar{G}(t|\bar{W}(t-))\Delta(t)dM(t)}{\bar{G}(t|X)},$$

where $\Delta(t) = I(C_r > t)$, and $\bar{G}(t \mid X)$ is censoring survival probability at time t. The performance of this estimating equation will be as good as the naive approach as discussed in subsection 2.1.1. $\bar{G}(t|\bar{W}(t-))$ is censoring survival probability conditional only on the covariates that are included into the model. Implementation of these estimating functions is relatively straight forward. We use standard software since S-plus routine `coxph()` can incorporate weights. We estimate the weights using the estimated censoring survival

Variable	Estimate	std.err.	Wald z	P-value
BEGAGE	-0.7055	0.0533	-13.25	0.000
CANDIDA	0.3794	0.1671	2.27	0.023
WTPCTA	-0.0052	0.0022	-2.39	0.017
SPUTMACT	-0.1373	0.0668	-2.06	0.040

Table 4: Estimated coefficients for the censoring intensity model.

probability obtained by selecting a model for censoring mechanism.

The first step in obtaining the desired estimating functions is to obtain a "good" model for the censoring mechanism. We assume multiplicative intensity model for the censoring counting process and initially consider all the covariates that we also considered when fitting Andersen-Gill for the intensity of exacerbations. The estimated model coefficients for the censoring mechanism are given in Table 4. Age at entry is the most important covariate saying that the older the patient is at entry, the less likely he or she is to be lost to follow up. This could simply be an artifact of our data set since the end of our study is reached once subject reaches age 14 and therefore, if a subject was lost to follow up, we do not observe it if it happened past age 14. Presence of Candida greatly increases the probability of being lost to follow up although it is not clear why this is the case. Presence of Candida IS not a significant covariate in the analysis of occurrence of exacerbations and a clinical explanation of this finding is not readily available. Increased weight for age percentile decreases the chance of censoring which indicates that the better the developmental status of a patient, the smaller the probability of censoring. However, the effect of weight is not very significant in clinical terms since the relative intensity for a 20 unit difference in weight for age percentile corresponds to only 0.9 relative intensity of censoring. Increased sputum productivity decreases the probability of censoring. Based on these results, there is a possibility that we have dependence between censoring and the counting process of interest conditional on the covariates included into model of interest. We see that not all covariates that are important for the censoring mechanism are included into our previously obtained intensity model of recurrent lung exacerbations. Therefore, we use the obtained censoring mechanism to estimate the censoring survival probability for every observation in our data set. Then we use the inverse of the estimated survival as the weights in recurrent exacerbations regression analyses. It is interesting to note that the estimated censoring survival probabilities in our data set range from 0.66 to 1. The regression coefficients estimated by the IPCW estimating function are given in Table 5. By comparing Tables 5 and 3, we see that the estimated coefficients or standard errors do not change noticeably when we employ weighting by the inverse of the probability of censoring. Therefore, even if the censoring due to loss to follow up is dependent, the effect of this dependence on the estimated coefficients is marginal. Both models yield the same clinical conclusions. FEV1 has a strong effect on the intensity of exacerbations with relative intensity 0.99. The higher the FEV1, the lower the intensity of exacerbations. Based on the estimated coefficient we see that the 10 unit difference in FEV1 corresponds to relative intensity 0.9, and the difference of 30 units to relative intensity 0.74. The

Variable	Estimate	std.err.	Wald z	P-value
SEX	0.0893	0.0252	3.30	0.0009
BEGAGE	0.4061	0.0147	27.68	0.0000
FEV1	-0.0105	0.0006	-16.20	0.0000
TOT1	0.8446	0.0370	22.70	0.0000
TOT2	1.3423	0.0441	29.37	0.0000
TOT3	1.6372	0.0512	31.15	0.0000
TOT4	1.8486	0.0593	27.36	0.0000
TOT5	2.1979	0.0661	32.63	0.0000
TOT6	2.1620	0.0753	25.35	0.0000
TOT7	2.4706	0.0825	29.92	0.0000
TOT8	2.5526	0.0609	33.20	0.0000
P.A.	0.2333	0.0349	6.33	0.0000
P.A. EVER	0.2736	0.0486	4.70	0.0000
B. CEPACIA	0.3716	0.0661	5.13	0.0000

Table 5: Estimated coefficients for the Andersen-Gill intensity model of recurrent lung exacerbations in the presence of possibly dependent censoring.

effect of the past of the process of interest on its intensity now is captured by covariate TOT that is present in the final model as a series of indicator variables. The intensity of exacerbations increases to the greatest extent after the first exacerbations has occurred although the higher the number of previous exacerbations, the higher the current intensity. The relationship between the past number of exacerbations with their current intensity is adjusted by BEGAGE that essentially tells us when we started observing the process and therefore gives different meanings to different values of TOT. Finally, as we expect, presence of microorganisms in respiratory cultures increases the intensity of exacerbations. Presence of B. Cepacia increases the intensity almost one and a half times while Pseudomonas presence has a notable effect in terms of increase not only based on the last culture results but also based on ever having had Pseudomonas detected in respiratory bacterial culture. A patient having previously had a culture positive for Pseudomonas as well as having the last culture positive, has 1.66 the intensity of a patient that has never had a culture positive for Pseudomonas.

While the inference drawn from the fitted models does not change depending on the estimating equation we use, we need to consider the weighted estimating equation approach in order to assess the possible impact of censoring. In this particular case, we found that the 8% of the population whose censoring times are assumed to follow the estimated conditional censoring distribution, do not have an impact on the inference we wish to draw. However, simply ignoring the possibility of dependent censoring may cause us to draw inference based on inconsistent parameter estimates in other applications.

5 Proportional rates model

We briefly reviewed the proportional rates models conditional on $\bar{Z}^*(t) \subset \bar{Z}(t)$ in the introduction and noted that it is suitable for a large number of recurrent events. The proportional rates model is very interesting since by not adjusting for the event history, it is producing interpretable regression coefficients for the baseline covariates. We will now provide a class of IPTW estimators for the proportional rates model given in (3) in the presence of dependent censoring.

Lin et al. (2000) proposed using the analogue of the Andersen-Gill partial likelihood estimating functions to obtain parameter estimates for the proportional rates model. The obtained estimates are only consistent and asymptotically normally distributed under the assumption that censoring only depends on the covariates entering the proportional rates model: i.e. $\lambda_C(t \mid \bar{X}(t)) = \lambda_C(t \mid \bar{Z}^*(t))$. In addition, they are inefficient, in general, even if the full data is observed. The reason for this is that partial likelihood is not the correct likelihood in the case of proportional rates.

Other models mentioned in the introduction suffer from similar problems since using the analogue of the partial likelihood estimating functions of the Andersen-Gill model is the proposed method for obtaining parameter estimates. Therefore these methods assume that the censoring mechanism does not depend on any covariates that are not already included into the model. This assumption becomes more questionable as the conditioning set decreases which is what the use of proportional rates models encourages.

The methods described above are readily applicable to the proportional rates model as well. In this model, one can use $D_h = \int h(t, \bar{Z}^*(t)) dM_r(t)$ as a class of full data estimating functions where $dM_r(t) \equiv dN(t) - E(dN(t) \mid \bar{Z}^*(t))$ and h is arbitrary. As in the full data intensity models, the desired set of estimating functions (which are not affected by the estimation procedure of the nuisance parameters) is a subset of this class of estimating functions. We can map these full data estimating functions into a class of observed data estimating functions with the same above presented mapping $U_G(. \mid D_h)$. In particular, applying our proposed choice for the index h of the full data estimating function, we get

$$U_G^r(Y \mid D) = \tag{18}$$
$$\int_0^\tau \left[\gamma^*(t) - \frac{E[\frac{I(C>t)}{\bar{G}(t|X)} \gamma^*(t) \bar{G}(t \mid \bar{Z}^*(t-)) Y(t) \exp(\beta \gamma^*(t))]}{E[\frac{I(C>t)}{\bar{G}(t|X)} \bar{G}(t \mid \bar{Z}^*(t-)) Y(t) \exp(\beta \gamma^*(t))]} \right] \frac{\bar{G}(t \mid \bar{Z}^*(t-)) \Delta(t)}{\bar{G}(t \mid X)} dM_r(t).$$

This yields simple to implement estimators which are at least as efficient as the "partial likelihood" based estimating functions used in Lin et al. (2000). These estimators remain consistent even if $\lambda_C(t \mid \bar{X}(t)) \neq \lambda_C(t \mid \bar{Z}^*(t))$ as long as the censoring mechanism is estimated consistently and the identifiability assumption (14) holds.

6 Discussion

We illustrated the substantial bias that can be introduced to the estimators of the unweighted estimating function derived from the observed data partial likelihood in the case of informative censoring and showed that the presented IPCW estimator is unbiased and performs much better compared to the naive estimator. Even though this estimator additionally requires the modeling of the censoring mechanism, the weights that utilize the censoring mechanism reduce to 1 in the case of independent censoring, hence the performance of the estimator is not effected by model of the censoring mechanism under independent censoring. Although there is no notable difference in the data analysis results we obtain using weighted or unweighted estimating functions, it is important to account for possibly dependent censoring. Simply said, we would not be able to assess the impact of censoring on the estimated coefficients had we not implemented the proposed approach. In addition, implementation of the proposed approach is simple and existing software can be used.

7 Appendix

Before the discussion of constructing the doubly robust estimators for the full data parameter of interest, we will define and introduce some notation and terminology. Let μ denote the parameter of interest in the full data model (i.e. regression parameter β in model (4)) and η denote the possible nuisance parameters in this model (i.e. λ_0 in model (4)). The nuisance tangent space is defined as the closure of the linear span of nuisance scores of one-dimensional sub-models for which the pathwise derivative of parameter of interest, μ, equals zero (e.g. see van der Laan and Robins (2002), p.55; Bickel et al. (1997) for the general theory of (nuisance) tangent spaces and pathwise derivatives). In particular, we will denote the orthogonal complement of the nuisance tangent space in the full data model of interest (i.e. given in (4)) by $T^{F,\perp}_{nuis}$. Let $L^2_0(P_{F_X,G})$ denote the Hilbert space of functions of Y with finite variance and mean zero and endowed with the covariance inner product $< f, g >_{P_{F_X,G}} = E_{P_{F_X,G}} f(Y)g(Y)$. The observed data estimating functions given in (13) are elements of this Hilbert space. Recall from Section 1 that, the CAR assumption on the censoring mechanism causes the observed data likelihood to factorize into a F_X- part and a $G-$part. We will refer to the F_X part as Q_X.

7.1 Orthogonalized observed data estimating functions

We will construct a doubly robust estimator for our full data parameter of interest β in model (4) using the general methodology of van der Laan and Robins (2002) (p.81). This general methodology requires the orthogonalization of an initial observed data estimating function $U_G(D(. \mid \mu, \eta))$ (i.e. given in (15)), that satisfies $E_{P_{F_X,G}}[U_G(D(. \mid \mu(F_X), \eta(F_X))) \mid X] = D(. \mid \mu(F_X), \eta(F_X)) \in T^{F,\perp}_{nuis}$, with respect to the T_{CAR}. Here, T_{CAR} is the nuisance tangent space for the censoring mechanism G in the observed data

model for $P_{F_X,G}$ only assuming CAR and is given by

$$T_{CAR} = T_{CAR}(P_{F_X,G}) = \{V(Y) : E(V(Y) \mid X) = 0\} \subset L_0^2(P_{F_X,G}),$$

where $V(Y)$ represents functions of the observed data random variable Y. The T_{CAR} orthogonalized estimating function is defined as

$$IC(Y \mid Q_X, G, D(. \mid \mu, \eta)) = U_G(D(. \mid \mu, \eta)) - IC_{CAR}(Y \mid Q_X, G, D(. \mid \mu, \eta)), \qquad (19)$$

where $IC_{CAR}(Y \mid Q_X, G, D(. \mid \mu, \eta)) \equiv \Pi(U_G(D(. \mid \mu, \eta)) \mid T_{CAR})$ is the projection of the initial estimating function $U_G(D(. \mid \mu, \eta))$ onto T_{CAR}.

This orthogonalized estimating function has the so called *double robustness property* (Robins et al. (2000); van der Laan and Robins (2002), p.81). The double robustness property allows misspecification of either the censoring mechanism $G(. \mid X)$ or the Q_X part of the full data distribution. Let Q_X^1 and $G_1 \in \mathcal{G}(CAR)$ be guesses of Q_X and $G(. \mid X)$, respectively. Then, we have

$$E_{Q_X,G} IC(Y \mid Q(X^1), G_1, D(. \mid \mu(F_X), \eta(Q_X^1))) = 0$$

if either $G_1 = G(. \mid X)$ and $G(. \mid X)$ satisfies the identifiability condition $\bar{G}(\tau \mid X) > \delta > 0, F_X - a.e.$ (given in (14)) or $Q_X^1 = Q_X$ and without any further assumptions on $G(. \mid X)$.

Moreover, the influence curve of μ using the estimating function (19) is given by

$$IC(\mu) = -\left[\frac{\partial}{\partial \mu} E_{P_{Q_X,G}} IC(Y \mid Q_X, G, D(\cdot \mid \mu, \eta))\right]^{-1} IC(Y \mid Q_X, G, D(\cdot \mid \mu, \eta)).$$

If we assume the correct model for G where it satisfies the identifiability assumption and an incorrect one for Q_X, then the resulting estimator is still consistent and asymptotically normal because the estimating function is still unbiased. However, $IC(\mu)$ and therefore the estimated variance based on it, are not correct although the resulting confidence intervals are conservative and can be used. For true influence curve see van der Laan and Robins (2002) (p.146). If the assumed model for G is incorrect and the model for Q_X is correct, the resulting estimator is consistent and asymptotically normal although $IC(\mu)$ is incorrect and bootstrap can be used to estimate the variance. Practical performance of a doubly robust estimator constructed using this methodology is illustrated by Yu and van der Laan (2002) in another data structure (longitudinal marginal structural models).

7.2 Orthogonalized estimating function for the marginal Anderson-Gill multiplicative intensity model

We now apply the above methodology to the IPCW estimator $U_G(. \mid D_h)$ (simply referred as $U_G(D)$ below) proposed for recurrent events data.

The projection of the $U_G(D)$ onto T_{CAR} equals (van der Laan and Robins (2002), Theorem 1.1, p.39),

$$\Pi(U_G(D) \mid T_{CAR}) = \int \left[E(U_G(D) \mid \bar{X}(u), C = u) - E(U_G(D) \mid \bar{X}(u), C > u)\right] \, dM_C(u),$$

where $dM_C(u) = dA(u) - \lambda_C(u)$ is a martingale with respect to the censoring process $A(t) = I(C \leq t)$, at time u. For $U_G(D)$ given in (13), we note that $E(U_G(D) \mid \bar{X}(u), C = u) = E(U_G(D) \mid \bar{X}(u), C > u)$ for $t < u$, and $E(U_G(D) \mid \bar{X}(u), C = u) = 0$ for $t \geq u$ so that the projection equals to

$$\Pi(U_G(D) \mid T_{CAR}) = -\int E\left[\int_u^\tau \frac{h^\star(t, \bar{W}(t-))\, dM(t)\Delta(t)}{\bar{G}(t \mid X)} \;\middle|\; \bar{X}(u), C > u\right] dM_C(u),$$

where $h^\star = h(t, \bar{W}(t-)) - g(h)$ as given in Section 2.1. The proposed doubly robust estimator is now the solution of the following estimating function:

$$\begin{aligned} IC(Y|\phi(Q_X, G), G, D(X|\mu, \lambda_0)) = \\ = \int_0^\tau \frac{h^\star(t, \bar{W}(t-))\, dM(t)\Delta(t)}{\bar{G}(t \mid X)} + \int_0^\tau \phi(Q_X, G)(u, \bar{X}(u))\, dM_C(u) \end{aligned}$$

where

$$\phi(Q_X, G) = E_{Q_X,G}\left[\int_u^\tau \frac{h^\star(t, \bar{W}(t-))dM(t)\Delta(t)}{\bar{G}(t \mid X)} \;\middle|\; \bar{X}(u), C > u\right].$$

Our estimating equation depends on $\phi(Q_X, G)$, G and λ_0 in addition to the parameter of interest β. As before, one can use Andersen-Gill multiplicative intensity model for the censoring mechanism to obtain an estimate of $\bar{G}(t|X)$, and use the estimator given in (16) for the baseline intensity λ_0. $\phi(Q_X, G)$ is the expectation given by the projections of $U_G(D)$ onto T_{CAR} and it also needs to be estimated. Since we are dealing with an integral, we can approximate it with a sum of estimated values obtained from a repeated measures regression (van der Laan and Robins (2002), p.201), this corresponds to directly estimating $\phi(Q_X, G)$, (i.e. without estimating Q_X and G components separately). An alternative method for estimating this nuisance parameter is by estimating Q_X and G and then estimating $\phi(Q_X, G)$ by monte carlo simulation method (van der Laan and Robins (2002), p.198). In this approach, one assumes a model for the complete full data generating distribution and estimate the model parameters by maximum likelihood estimation. Then the conditional expectations of the form $\phi(Q_X, G)$ under this fitted model is estimated by monte-carlo simulation. As we noted previously, if our model for $\phi(Q_X, G)$ is incorrect, as long as we model the censoring mechanism correctly, the resulting estimates are consistent and asymptotically normal.

Since we often can not rely on assuming the correct model for Q_X, one might be concerned with the possible loss in estimating equation efficiency that might occur when we subtract an estimate of the projection onto T_{CAR} from $U_G(D)$. In order to ensure increase in efficiency relative to an initial $U_G(D)$, assuming a correctly specified model for G, one can use the following estimating equation (Robins and Rotnitzky (1992))

$$IC(Y \mid Q, G, D(\cdot \mid \mu, \eta), c_{nu}) = U_G(D) - c_{nu}\Pi(U_G(D)|T_{CAR}),$$

where the matrix c_{nu} equals (using a shorthand notation IC_{CAR} for $\Pi(U_G(D)|T_{CAR})$).

$$c_{nu} = E_{P_{F_X,G}}[U_G(D)IC_{CAR}]E_{P_{F_X,G}}[IC_{CAR}IC_{CAR}^t]^{-1},$$

and can be estimated with

$$c_{nu,n} =< \hat{U}_G(D), \hat{IC}_{CAR} >_n < \hat{IC}_{CAR}, \hat{IC}_{CAR}^t >_n^{-1} .$$

Here $\langle h, g \rangle_n \equiv 1/n \sum_{i=1}^n h(Y_i)g(Y_i)$. If $\hat{IC}_{CAR}$ consistently estimates $\Pi(U_G(D)|T_{CAR})$, then $c_{nu,n}$ consistently estimates the identity matrix. Therefore c_n can also be used as a method for selecting the best fit for IC_{CAR} among a number of candidates $\hat{IC}_{CAR}$. We refer to van der Laan and Robins (2002) (p.142) for a more detailed treatment of this extension.

References

P.K. Andersen, O. Borgan, R.D. Gill, and N. Keiding. *Statistical Models Based on Counting Processes*. Springer-Verlag New York, 1993.

P.K. Andersen and R.D. Gill. Cox's regression for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120, 1982.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J. Wellner. *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, 1997.

R.D. Gill. Understanding Cox's regression model: A martingale approach. *Journal of the American Statistical Association*, 79(386):441–447, June 1984.

R.D. Gill, M.J. van der Laan, and J.M. Robins. Coarsening at random: characterizations, conjectures and counter-examples. In D.Y. Lin and T.R. Fleming, editors, *Proceedings of the First Seattle Symposium in Biostatistics*, pages 255–94, New York, 1997. Springer Verlag.

D.F. Heitjan and D.B. Rubin. Ignorability and coarse data. *Annals of statistics*, 19(4): 2244–2253, December 1991.

M. Jacobsen and N. Keiding. Coarsening at random in general sample spaces and random censoring in continuous time. *Annals of Statistics*, 23:774–86, 1995.

J.F. Lawless. The analysis of recurrent events for multiple subjects. *Applied statistics - Journal of the Royal Statistical Society Series C*, 44(4):487–498, 1995.

J.F. Lawless and C. Nadeau. Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37(2):158–168, May 1995.

J.F. Lawless, C. Nadeau, and R.J. Cook. Analysis of mean and rate functions for recurrent events. In *Proceedings of the First Seattle Symposium in Biostatistics*, volume 123 of *Lecture notes in statistics (Springer-Verlag)*, pages 37–49. New York : Springer, 1997.

D.Y. Lin, L.J. Wei, I. Yang, and Z. Ying. Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society series B - Statistical Methodology*, 62(PT4):711–730, 2000.

W.J. Morgan, S.M. Butler, C.A. Johnson, A.A. Colin, S.C. FitzSimmons, D.E. Geller, M.W. Konstan, M.J. Light, H.R. Rabin, W.E. Regelmann, D.V. Schidlow, D.C. Stokes, M.E. Wohl, H. Kaplowitz, M.M. Wyatt, and S. Stryker. Epidemiologic Study of Cystic Fibrosis: design and implementation of a prospective, multicenter, observational study of patients with cystic fibrosis in the U.S. and Canada. *Pediatric Pulmonology*, 28(4): 231–241, October 1999.

D. Oakes. Frailty models for multiple event times. In J.P. Klein and P.K. Goel, editors, *Survival Analysis: State of the Art*, pages 371–379. Kluwer Academic Publishers, Netherlands, 1992.

M.S. Pepe and J. Cai. Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association*, 88(423):811–820, September 1993.

Y. Ritov and J.A. Wellner. Censoring, martingales and the Cox model. *Contemporary Mathematics*, 80:191–219, 1988.

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

J.M. Robins and A. Rotnitzky. Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology*, Methodological issues. Bikhäuser, 1992.

J.M. Robins, A. Rotnitzky, and M. van der Laan. Comment on "On Profile Likelihood" by S.A. Murphy and A.W . van der Vaart. *Journal of the American Statistical Association – Theory and Methods*, 450:431–435, 2000.

M.J. van der Laan and J.M. Robins. Unified methods for censored longitudinal data and causality. To be published by Springer, 2002.

L.J. Wei, D.Y. Lin, and L. Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073, December 1989.

Z. Yu and M.J. van der Laan. Doubly robust estimators for longitudinal marginal structural models. To be submitted, 2002.