University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2007 *Paper* 218

Regression Analysis of a Disease Onset Distribution Using Diagnosis Data

Jessica G. Young[*] Nicholas P. Jewell[†]

Steven J. Samuels[‡]

[*] University of California, Berkeley, jyoung@hsph.harvard.edu

[†] Division of Biostatistics, School of Public Health, University of California, Berkeley, jewell@berkeley.edu

[‡] State University of New York at Albany, ssamuels@albany.edu

Regression Analysis of a Disease Onset Distribution Using Diagnosis Data

Jessica G. Young, Nicholas P. Jewell, and Steven J. Samuels

Abstract

We consider methods for estimating the effect of a covariate on a disease onset distribution when the observed data structure consists of right-censored data on diagnosis times and current status data on onset times amongst individuals who have not yet been diagnosed. Dunson and Baird (2001) approached this problem using maximum likelihood, under the assumption that the ratio of the diagnosis and onset distributions is monotonic non-decreasing. As an alternative, we propose a two-step estimator, an extension of the approach of van der Laan, Jewell and Petersen (1997) in the single sample setting, that is computationally much simpler and requires no assumptions on this ratio. A simulation study is performed comparing estimates obtained from these two approaches, as well as that from a standard current status analysis that ignores diagnosis data. Results indicate that the Dunson and Baird estimator outperforms the two-step estimator when the monotonicity assumption holds, but the reverse is true when the assumption fails. The simple current status estimator loses only a small amount of precision in comparison to the two-step procedure but requires monitoring time information for all individuals. In the data that motivated this work, a study of uterine fibroids and chemical exposure to dioxin, the monotonicity assumption is seen to fail. Here, the two-step and current status estimators both show no significant association between the level of dioxin exposure and the hazard for onset of uterine fibroids; the two-step estimator of the relative hazard associated with increasing levels of exposure has the least estimated variance amongst the three estimators considered.

1 Introduction

There are many applications in epidemiology where the research question of interest involves the effect of some exposure on time to onset of a given disease. For example, consider data collected from the Seveso Women's Health Study (SWHS), where researchers were interested in estimating the effect of dioxin exposure on time to onset (age) of uterine fibroids in women living in Seveso, Italy during a chemical explosion in 1976. As the exact *age of onset* is unobservable for this disease, study data consisted of right-censored data on *age of diagnosis* of fibroids collected via questionnaire in 1996. In addition, uterine ultrasounds were given at this time to assess the presence or absence of a fibroid for individuals with no prior diagnosis.

Assuming the disease under study is irreversible and is detectable during a preclinical latency phase, we can define this observed data structure more generally as n independent copies of

$$\{C \wedge T_2, \Delta_1 = I(T_1 < C), \Delta_2 = I(T_2 < C), Z\}, \tag{1}$$

where T_1 is time to disease onset, T_2 is time to diagnosis, C is a random screening time assumed independent of (T_1,T_2) and Z is a k-dimensional set of fixed covariates. Further define F_1 and F_2, and S_1 and S_2, as the distribution functions and survival functions for T_1 and T_2, respectively.

Based on this observed data structure, three types of observations are possible:

(i) Both onset and diagnosis have occurred prior to the screening, and the time of diagnosis is known exactly ($T_1 < T_2 < C, \Delta_1 = \Delta_2 = 1$);

(ii) Onset, but not diagnosis, has occurred prior to the screening ($T_1 < C < T_2, \Delta_1 = 1, \Delta_2 = 0$);

(iii) Neither onset nor diagnosis has occurred prior to screening ($C < T_1 < T_2, \Delta_1 = \Delta_2 = 0$).

Based on n independent copies of the observations (1), we can write the log likelihood of the data, conditional on C and Z, as follows

$$\begin{aligned} l(F_1, F_2|C, Z) &= \sum_{i=1}^{n} \delta_{2i} \log f_2(t_{2i}|z_i) + \delta_{1i}(1 - \delta_{2i}) \log\{S_2(c_i|z_i) - S_1(c_i|z_i)\} \\ &+ (1 - \delta_{1i}) \log S_1(c_i|z_i), \end{aligned} \tag{2}$$

where δ_{1i}, δ_{2i}, t_{2i}, z_i, and c_i are the i^{th} observed values of Δ_1, Δ_2, T_2, Z and C, respectively.

As the likelihood only involves the marginal distributions of T_1 and T_2, given Z, these are the only identifiable aspects of the conditional joint distribution, F, of (T_1, T_2). In addition, the implicit constraint underlying the likelihood (2), that $\text{pr}(T_1 < T_2|Z) = 1$, translates simply to the explicit constraint on the marginals that $F_1(t|Z) > F_2(t|Z)$ for any Z. This follows since for any pair of marginals (F_1, F_2) with $F_1 > F_2$, there exists a bivariate distribution with F_1 and F_2 as marginals and $\text{pr}(T_1 < T_2) = 1$. In the following, we thus consider regression models based solely on the marginal distributions F_1 and F_2.

Our primary interest is in how the covariates Z affect the distribution of T_1. The methods we present can be extended to a variety of regression models, but here we focus on a proportional hazards model for T_1 and, specifically, the relative hazard coefficients ϕ in

$$S_1(t|Z) = S_{01}(t)^{\exp(\phi Z)}. \tag{3}$$

Many authors have discussed nonparametric estimation of F_1 in the absence of covariates with this data structure (van der Laan et al, 1997; Turnbull and Mitchell, 1984; Kodell, Shaw and Johnson, 1982; Dinse and Lagakos, 1982). Dunson and Baird (2001) specifically address estimation of ϕ, based on the model (3), with an application to US national data on the pre-menopausal incidence of uterine fibroids. Defining $Q(t|Z) = \text{pr}(T_2 < t|T_1 < t, Z)$, equivalent to the ratio of the distribution functions $\frac{F_2(t|Z)}{F_1(t|Z)}$ (and, thus, $F_2(t|Z) = Q(t|Z)F_1(t|Z)$), they reparametrized the likelihood such that (2) becomes

$$\begin{aligned} l(F_1, Q|C, Z) &= \sum_{i=1}^{n} \delta_{2i} \log\{Q(t_{2i}|z_i) f_1(t_{2i}|z_i) + q(t_{2i}|z_i) F_1(t_{2i}|z_i)\} \\ &+ \delta_{1i}(1-\delta_{2i}) \log F_1(c_i|z_i) + \delta_{1i}(1-\delta_{2i}) \log\{1 - Q(c_i|z_i)\} \\ &+ (1-\delta_{1i}) \log S_1(c_i|z_i), \end{aligned} \tag{4}$$

where $q(t|z)$ is the derivative of $Q(t|z)$ with respect to t, with the constraint that $Q(t|Z)F_1(t|Z)$ is a distribution function for all Z. They avoided this active constraint by assuming $Q(t|Z)$ itself is a distribution function, finally modeling both $Q(t|Z)$ and $F_1(t|Z)$ as proportional odds models.

Dunson and Baird approached maximum likelihood estimation of regression parameters of the proportional odds model for F_1 and Q by flexibly parametrizing both $S_{01}(t) = 1 - F_{01}(t) \equiv 1 - F_1(t|Z = 0)$ and $Q_0^*(t) = 1 - Q(t|Z = 0)$. Specifically, $\frac{F_{01}(t)}{1-F_{01}(t)}$ and $\frac{Q_0(t)}{1-Q_0(t)}$ (i.e., the baseline odds functions) were both modeled as piecewise linear with nine breakpoints and parametrized in terms of the slopes of this piecewise linear function. To ensure the required monotonicity, the slopes were constrained to be non-negative.

The Dunson and Baird approach guarantees estimates of F_1 and F_2 with the desired stochastic ordering and, in theory, gives an efficient estimate of ϕ if both regression models and the assumption regarding $Q(t|Z)$ are correct. However, in practice the high-dimensional estimation is computationally intensive and the theoretical properties of the resulting estimators come into question when the dimension of Z is even moderate because of the large number of necessary parameters used for the baseline survival functions. Further, their approach requires the assumption of non-decreasing Q. Dunson and Baird argue in support of the assumption of non-decreasing $\frac{F_2(t|Z)}{F_1(t|Z)}$ for most chronic diseases, with t representing age, by claiming that "the proportion of diseased individuals that have been diagnosed is unlikely to decrease with age". However, there are many situations where this assumption may not hold over the entire ranges of t and Z ; fortunately, we show below that the adequacy of the assumption can be examined using the available data.

Dunson and Baird note that their methods can be extended to proportional hazards for F_1, and for Q. When using the Dunson and Baird approach, we assume such proportional hazards models throughout, namely, for $F_1 = 1 - S_1$ the model given in (3), and for $Q^* = 1 - Q$ the model

$$Q^*(t|Z) = Q_0^*(t)^{\exp(\psi Z)}. \tag{5}$$

Approaches for estimating ϕ in the context of the observed data structure (1) without restrictive assumptions on parameters of little interest such as Q are desirable. Ideally, ϕ would be estimated in this regression setting with a full (semiparametric) maximum likelihood estimator based on the likelihood (2) with appropriate specification of the marginal distribution functions, as was done in the nonparametric single sample setting by Turnbull and Mitchell (1984). However, in the regression setting it is difficult to describe an appropriate joint distribution for T_1 and T_2 that yields appropriate marginal regression models of interest, such as (3), and simultaneously satisfies the ordering constraint $F_1(t|Z) > F_2(t|Z)$ for all t, Z over the full range of possible regression coefficients. In addition, specification of a joint distribution may require description of the unidentifiable dependence structure between T_1 and T_2. Finally, even if this issue is ignored, there are substantial computational issues involved in maximizing (2) over a very high dimensional parameter space, assuming that the baseline components of the marginal regression models are described nonparametrically.

The constraint, $F_1(t|Z) > F_2(t|Z)$ for all t, Z, links estimation of $F_1(t|Z)$ and $F_2(t|Z)$ so that separate maximizations of (2) with regard to $F_1(t|Z)$ and $F_2(t|Z)$ risk the possibility of obtaining estimates with $F_1(t|Z) < F_2(t|Z)$ for some t, Z. However, if estimation of these distributions is not of primary interest, one might consider approaches that avoid full maximum likelihood estimation to focus more directly on the parameter of interest, ϕ, and make use of existing algorithms to make computation much simpler in practice. In §2 we propose such an approach, extending the two-step approach for estimation of F_1 proposed by van der Laan et al. (1997) for the single sample case to the semi-parametric regression setting.

Intuitively, the observed data structure (1), which contains observed values of T_2 in addition to coarsened or current status data (Jewell and van der Laan, 2004) on T_1, would provide more exact information about T_1 than pure current status data alone. One might assume, in turn, that this would allow for more efficient estimation of ϕ. van der Laan et al. (1997) showed that in the fully nonparametric single sample setting, estimators of $S_1(t)$ based solely on current status data on T_1, ignoring information on T_2, are often less efficient than alternatives. Note that in addition to ignoring information on T_2, pure current status data differs from (1) in that C must be observed for all n individuals, not only for those with $T_2 > C$. When this is the case, straightforward methods and algorithms for obtaining an estimator of ϕ are well established (Shiboski, 1998).

In §3, we describe a simulation study comparing the performance of estimators of ϕ based on the Dunson and Baird, two-step and pure current status approaches. In §4, we apply these approaches to the SWHS noted above.

2 Two-step approach for the estimation of ϕ

The following is a modification of the approach proposed by van der Laan et al. (1997) in the nonparametric single sample setting to the semi-parametric regression setting. The approach avoids the pitfalls of high-dimensional full maximum likelihood and the restrictive assumptions of Dunson and Baird (2001). Moreover, it can be easily applied using existing algorithms for proportional hazards regression analysis of right-censored and current status data.

The two-steps of the estimation algorithm can be stated as follows (note proportional hazards is assumed here for simplicity but this assumption is not necessary):

(i) Use the data, $\{c_i \wedge t_{2i}, \delta_{2i}, z_i : i = 1, \ldots n\}$, to estimate $S_2(c_i|z_i)$ using standard proportional hazards regression methodology;

(ii) Estimate the parameter of primary interest ϕ in $S_1(t|Z)$, arising from the proportional hazards model (3), using the data, $\{c_i, \delta_{1i}, z_i : i = 1, \ldots n\}$ only for those individuals for whom $\delta_{2i} = 0$; this is achieved by modifying an algorithm for proportional hazards regression for standard current status data and applying it to the constructed outcome $S_2(C|Z)(1 - \Delta_1)$ (based on the first step) and covariates Z.

The rest of this section motivates this approach and describes the proposed algorithm in more detail.

As noted by van der Laan et al. (1997), $R(c) = \frac{S_1(c)}{S_2(c)} = E\{1 - \Delta_1 | C, Z, T_2 > C\}$. It follows that

$$S_1(C|Z) = E\{S_2(C|Z)(1 - \Delta_1)|C, Z, T_2 > C\}.$$

This suggests that to estimate S_1, we perform a monotonic regression of $S_2(C|Z)(1-\Delta_1)$ on C and Z, amongst individuals with $T_2 > C$ (that is, $\Delta_2 = 0$). Specifically, the proportional hazards model (3) for $S_1(t|Z)$ leads to a generalized additive regression model for the unobserved random variable $d_1 = S_2(C|Z)(1 - \Delta_1)$ with the $\log\{-\log(.)\}$ link function for the mean, and regression function $\Lambda(C) + \phi Z$, where Λ is an arbitrary increasing function, determined by the baseline survivor function, $S_{01}(t)$, with the property that $\Lambda(C) \to \infty$ as $C \to \infty$ and

$\Lambda(C) \to -\infty$ as $C \to 0$. This follows since $\log\{-\log[E(d_1|C,Z)]\} = \log\{-\log[S_{01}(C)]\} + \phi Z$. Although d_1 is not observed, it can be estimated using $\hat{S}_2(c_i|z_i)(1-\delta_{1i})$, for the ith individual with $\delta_{2i} = 0$, using an estimator for $\hat{S}_2(c_i|z_i)$. Many flexible models for $S_2(t|Z)$ could be considered, but here, for simplicity, we assume a proportional hazards model for T_2; that is,

$$S_2(t|Z) = S_{02}(t)^{\exp(\beta Z)}, \tag{6}$$

with estimators of S_{02} and β easily obtained by standard methods for right-censored data (Cox, 1972). Note that alternative regression models for T_1 and Z to proportional hazards can be accommodated in this algorithm by merely choosing a different link function in the generalized additive model for d_1.

Note, for fixed C, Z,

$$\begin{aligned} \mathrm{var}\{S_2(C|Z)(1-\Delta_1)|C,Z,T_2 > C\} &= {S_2}^2(C|Z)\mathrm{var}(1-\Delta_1|C,Z,T_2 > C) \\ &= {S_2}^2(C|Z)R(C|Z)\{1-R(C|Z)\}. \end{aligned} \tag{7}$$

This suggests weighting the regression fit of $\hat{S}_2(C|Z)(1-\Delta_1)$ on C and Z with weights inversely proportional to ${S_2}^2(C|Z)R(C|Z)\{1-R(C|Z)\}$. These weights require knowledge of $R(C|Z)$ and thus $S_1(C|Z)$, and so must be iteratively updated along with estimation of $S_1(C|Z)$, described in additional detail below. A simpler weighting scheme employs ${S_2}^2(C|Z)$ instead of (7), as in van der Laan et al. (1997) in the single sample setting, although this choice of weights may cause some loss in precision in the semi-parametric setting.

Once an estimate of $S_2(C|Z)$ is obtained, the regression coefficients ϕ, as well as $\Lambda(C)$, can be estimated by modifying Shiboski's (1998) algorithm for regression with current status data. The modified algorithm is applied to the $\log\{-\log(.)\}$ regression of $\hat{d}_1$ on C and Z as noted (that is additive in C and Z, monotonic in C and linear in Z). The reader is referred to Shiboski (1998) for a detailed description of this algorithm, which is equivalent to the 'local scoring' procedure used to estimate generalized additive models (Hastie and Tibshirani, 1990). However, briefly, the algorithm consists of an outer loop where an adjusted dependent variable and weight are calculated based on the mean (μ) of the random variable of interest and an associated link function. In the case of proportional hazards, the appropriate link function is $\log\{-\log(1-\mu)\}$. This outer loop is followed by an inner backfitting loop, which alternates between estimating ϕ via weighted least squares and estimating $\Lambda(C)$ via the weighted 'Pool Adjacent Violators Algorithm' (Barlow et al., 1972).

Only minor modifications of Shiboski's (1998) algorithm are necessary to implement the two-step approach. These include: (i) the data to be passed to the former is a simple current status observation (e.g. δ_1) which has mean $\mu|Z = F_1(c|Z)$, whereas data to be passed to the latter is of the form $\hat{S}_2(c|z)(1-\delta_1)$, which, as shown above, has (approximate) mean $\mu|Z = S_1(c|Z)$; (ii) the appropriate link function in the proportional hazards case for the latter is, therefore, no longer $\log\{-\log(1-\mu)\}$ but, $\log\{-\log(\mu)\}$ which, accordingly, modifies the weights in the outer loop of the 'local scoring' procedure; (iii) these weights are further modified to accommodate the fact that the variance of the data is not that of a simple current status observation (e.g. $\text{var}(\Delta_1) = F_1(C)\{1-F_1(C)\}$), but rather that shown in (7); and (iv) the original algorithm is equivalent to maximizing the likelihood for standard current status data, whereas the latter involves only minimizing a weighted squared error - thus convergence criteria are modified from being based on changes in the deviance in the former to changes in the parameters themselves in the latter.

It is expected that the two-step estimator of ϕ may lose some efficiency over maximizing the full likelihood jointly as in Dunson and Baird (2001) when their monotonicity assumption regarding $Q(t|Z)$ is correct. However, when this assumption does not hold, the full likelihood approach is likely to result in biased effect estimators. To check the monotonicity assumption, $Q(t|Z)$ can be plotted from the estimates of F_1 and F_2 provided by the two-step approach.

A potential disadvantage of the two-step approach is that the constraint $\hat{S}_1(t|Z) < \hat{S}_2(t|Z)$ might be violated for some t and Z; however, this may be less likely to occur than in the case of the pure current status approach to estimating $\hat{S}_1$ since the constructed observations $\hat{S}_2(c_i|z_i)(1-\delta_{1i})$ in the regression model for $S_1(t|Z)$ are always less than or equal to $\hat{S}_2(t|Z)$.

If the monitoring time C is observed for all individuals, it is known that the estimator of ϕ in (3) is asymptotically Normal when maximum likelihood estimation is based on pure current status data on T_1 (Huang, 1996). Huang (1996) discusses methods for estimation of the limiting variance. In addition, the associated estimator of (the nuisance parameter) S_{01} is consistent, but converges only at rate $n^{-1/3}$ as compared to the standard $n^{-1/2}$ rate. In principal, the Dunson and Baird (2001) maximum likelihood estimator follows standard theory, although variance estimation is complicated by the high-dimensional form of the parametrization, and the dependency on the monotonic assumption regarding Q. Dunson and Baird (2001) suggest the use of profile likelihood confidence intervals to deal with the first of these issues.

Asymptotic theory regarding the two-step estimator of ϕ remains to be fully articulated, although the work of van der Laan et al. (1997) strongly suggests that the estimator converges at rate $n^{-1/2}$ and is asymptotically Normal assuming that, as the sample size increases, monitoring times are selected to allow the standard estimator of $S_2(c|z)$ to converge consistently at rate $n^{-1/2}$. Formal proofs of such asymptotic results for semiparametric regression models with current status data are complicated by the non-standard rate of convergence of estimators of the nuisance parameter, S_{01}, as noted above. However, general techniques based on locally efficient estimating equations with incomplete data (van der Laan and Robins, 2003) support the validity of this conjecture—see also Andrews, van der Laan and Robins (2005). Note that, in the case where the distribution of F_2 is entirely supported at ∞ (or practically, at large values), the two-step and the current status estimators coincide since all data is then of the current status form—in this case, the asymptotic results for the two-step estimator therefore follow immediately from Huang (1996).

Despite this support for the standard convergence of the two-step estimator of ϕ, its influence curve is unlikely to provide the basis for an effective 'plug-in' estimator of its variance as demonstrated in a related current status data estimation problem (Jewell, van der Laan and Lei, 2005). However, in the context of bivariate current status data, Jewell et al. (2005) note that the simple bootstrap provides a potential way of obtaining an estimate of the variance of smooth functional estimators and is more sensitive to second-order asymptotic properties. Note that Ma and Kosorok (2005) suggest the possible use of a weighted bootstrap in this kind of problem with non-standard rates of convergence for estimators of the nuisance parameters, and provide a theoretical justification. The weighted bootstrap has also been used by Ma and Kosorok (2006) and Strawderman (2006) in other similar applications. An alternative approach, that may be less computationally intensive, involves sampling from the posterior of the profile likelihood for ϕ (Lee, Kosorok and Fine, 2005). We evaluate the accuracy of the simple and weighted bootstrap approaches to estimating the variance of the two-step estimator of ϕ in §3.

3 Simulations

In this section, we compare estimates of the regression coefficient ϕ of the marginal proportional hazards model (3) for S_1 based on the Dunson and Baird, two-step and current status methods. For simplicity, we focus on the case of a single covariate Z. Data were simulated corresponding to the observed data structure (1), as opposed to using an assumed joint distribution for (T_1,T_2), as it is difficult to formulate a joint distribution function that integrates to (3) and (6) for all Z. Specifically, for the i^{th} observation, the covariate value z_i

was first generated from a Uniform$\{0,1\}$; the outcome t_{2i} was then simulated based on the model (6), for specific choices of S_{02} and β, and corresponding c_i generated independently from a fixed monitoring time distribution. The observed value of δ_{2i} follows from comparison of t_{2i} and c_i. For specified values of S_{01} and ϕ, each δ_{1i} was generated from a Bernoulli(p), where $p = 1 - \frac{S_1(c_i|z_i)}{S_2(c_i|z_i)}$, amongst observations with $\delta_{2i} = 0$ only; here $S_1(c|Z)$ is determined by (3).

Four simulations were performed with the following specifications: for simulations 1 and 2, the baseline distributions, S_{01} and S_{02} were defined such that $0.014T_1 \sim$ Weibull$(2.0, 2.5)$ and $0.014T_2 \sim$ Weibull$(1.99, 4.5)$. For simulations 3 and 4, the baseline distributions were modified so that $T_1 \sim$ Weibull$(-10.2, 2.5)$ and $T_2 \sim$ Weibull$(-10.3, 2.4)$. In all cases, C was generated independently from a Uniform$(20, 60)$ distribution; this range for C was motivated by the age range of screening times in the SWHS data (see §4). The individual simulations differed based on the population values of ϕ and β: the coefficients were selected to be $\phi = \beta = -0.4$ for simulations 1 and 3, and $\phi = \beta = 0$ for simulations 2 and 4.

Note that the simplicity of the factor Z allows all these choices of $S_{01}, S_{02}, \phi, \beta$, so that the ordering constraint $F_1(t|Z) > F_2(t|Z)$ is satisfied for $t \in (20, 60)$. On the other hand, the simulation parameters lead to differing shapes for the function $Q(t|Z)$ for $Z \in \{0,1\}$ and over this range of t. Specifically, $Q(t|Z)$ is monotonic non-decreasing in both simulations 1 and 2, but is non-monotonic for simulations 3 and 4, thus violating the Dunson and Baird (2001) assumption. In the latter two scenarios the shape of $Q(t|Z = 0)$ is approximately quadratic, dropping from a value close to 0.68 at $t = 20$ to a minimum of 0.665 around $t = 32$, and then rising to approximately 0.72 at $t = 60$.

The pure current status estimator of ϕ was obtained using the regression techniques for current status data described in Shiboski (1998) using an R package provided by the author.

The estimator of ϕ based on the Dunson and Baird approach was obtained as described in their paper (Dunson and Baird, 2001) with the slight modification of using proportional hazards models in place of proportional odds models for $F_1(t|Z)$ and $Q(t|Z)$. This involved assuming piecewise linear models for $-\log S_{01}(t)$ and $-\log Q_0^*(t)$ (i.e. the baseline cumulative hazards), both with nine breakpoints at $t \in \{7, 14, 21, 28, 35, 42, 49, 56, 63\}$.

Five hundred simulations, each based on sample size $n = 500$, were evaluated under the four simulation scenarios. Table 1 presents a comparison of the three estimation procedures in terms of bias, variance and mean squared error. Note that a relative efficiency of less than one in Table 1 reflects superior performance of the two-step estimator in the relevant comparison, and vice-versa. To give a general impression of the simulated data, the fraction of observations with $\Delta_2 = 1$ is approximately 41%, 45%, 19%, and 22% for simulations 1,2, 3, and 4, respectively; amongst those individuals with $\Delta_2 = 0$, the fraction with $\Delta_1 = 1$ is 48%, 54%, 11%, and 13%, for simulations 1,2,3, and 4, respectively. TABLE 1 ABOUT HERE

When Q is monotonic non-decreasing, the Dunson and Baird (2001) estimator slightly outperforms the other two estimators as might be expected since it is intended to be full maximum likelihood. Notably, the method does not seem affected by the fact that, in the simulated data, the regression model we used for Q is not correct. On the other hand, when Q is non-monotonic, the Dunson and Baird (2001) estimator suffers from considerable bias. Even from these limited simulations, the validity of the assumption on Q seems to be crucial in recommending use of this estimator.

The two-step and current status estimators are essentially equivalent when Q is decreasing, with a moderate advantage for the two-step method when this assumption is violated. The fact that the two-step estimator is not universally superior is somewhat counter-intuitive given that the current status method ignores data on T_2. This result, however, is in line with those of van der Laan and Jewell (2003) for smooth functionals in the absence of covariates. Recall that the current status method is only feasible when monitoring times are observed for all subjects.

Table 2 presents, for each of the four simulations, the variance of $\hat{\phi}$, along with the median of the simple and weighted bootstrapped estimates of this variance across the 500 simulations. The simple bootstrap variance estimates are based on 500 replicates of simple, unweighted, samples of size n (with replacement). The weighted bootstrapped estimates are also based on 500 replicates with weights selected from a unit exponential distribution for each replicate (see Ma and Kosorok (2005) for details). In all four simulations, the median of the variance estimators for both the simple and weighted bootstrap is close to the actual variance; however the simple bootstrap appears to perform better in practice. TABLE 2 ABOUT HERE

In the simulations of Table 1, the form of $S_2(t|Z)$ was correctly modeled in the implementation of both the two-step and current status approaches. The simpler current status approach requires no assumptions about the distribution of T_2 and is, thus, robust against misspecification of S_2. In the case of the two-step method, any parametric or semi-parametric model can be used to estimate $S_2(t|Z)$ in practice (not just proportional hazards). Further, data adaptive methods provide another approach to avoiding misspecification of this nuisance parameter. To assess the robustness of the two-step method against misspecification of S_2, simulations 1 through 4 were repeated with the true S_2 depending not only on Z but also on another covariate, W. Thus, instead of (6), the correct form of S_2 in this case is $S_{02}^{\exp(\beta Z+\gamma W)}$. The two-step method was then applied to this alternative simulated data including only Z in the estimation of S_2 and erroneously omitting the important covariate, W (generated from a Uniform$\{0,1\}$). As shown in Table 3, misspecification of S_2 in this scenario does not substantially alter the properties of the two-step estimator of ϕ compared with those when S_2 is correctly specified (see Table 1). TABLE 3 ABOUT HERE

4 Application

To present an application of the three methods compared in §3, we estimated the effect of exposure to the compound 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) on time to onset of uterine fibroids—non-cancerous tumors of the uterus often associated with reproductive dysfunction—based on data from the Seveso Women's Health Study (SWHS). SWHS is a retrospective cohort study of women living in Seveso, Italy as of July 10, 1976, when a chemical explosion exposed residents to the highest known levels of TCDD in a human population (Eskenazi et al., 2000).

Study participants had their serum TCDD levels ascertained soon after the explosion and were then followed up approximately twenty years later, when they were given an in-depth interview regarding their medical history, including any history of a fibroids diagnosis. At this same time, participants were offered ultrasounds to detect sub-clinical onset of uterine fibroids. Using the notation in (1), C here represents the age at ultrasound/interview, T_2 the age at fibroids diagnosis based on interview/medical records, T_1 the age at fibroids onset (always unobserved), Δ_1 the indicator of whether a fibroid was found at ultrasound, and Δ_2 the indicator of whether any history of a fibroids diagnosis was reported in the interview/medical records. Z is defined as the $\log_{10}$ serum TCDD level collected following the explosion. The analysis consisted of 956 women between 0 and 40 years of age at the time of the explosion, with no prior diagnosis of fibroids at this time. A more detailed description of the study procedures, including a more in-depth data analysis, is presented in Eskenazi et al. (2007).

A disadvantage of these data is that Δ_1 is missing in a substantial proportion of the sample. Specifically, 209 women with no prior diagnosis of fibroids ($\Delta_2 = 0$) were not offered or refused an ultrasound, and thus were missing the current status indicator, Δ_1. How these observations were dealt with in estimating $S_1(t|Z)$ in the analyses described below varies by approach. In analyses based on the Dunson and Baird (2001) method, this issue was addressed in the manner described in their paper by adding an additional term, $S_2(c_i|z_i)$, to the likelihood (4) for the i^{th} individual missing current status data and no prior diagnosis of fibroids. For the two-step approach, these 209 observations were included in the estimation of $S_2(t|Z)$ in step 1, but necessarily excluded from step 2, which requires knowledge of Δ_1. Finally, as these incomplete observations only contain information on T_2, they are entirely excluded from the pure current status approach, which does not use T_2 data at all.

Of the 956 women in the total analysis sample, 763 (80%) had not been diagnosed with fibroids by their screening age, c_i. Of these 763 observations, 554 additionally had non-missing values for Δ_1; of these, 58 (10.5%) had fibroids detected at ultrasound ($\Delta_1 = 1$). Therefore, step 1 of the two-step analysis is based on all 956 observations, while step 2 is based on this subgroup of 554 with $\Delta_2 = 0$ and nonmissing Δ_1. The screening ages, C, ranged from 20 to 60 years and TCDD levels (original scale) ranged between 2.5 and 56000.0 parts per trillion (ppt). Note that the observed distribution of (Δ_1, Δ_2) most closely resembles simulations 3 and 4 of §3.

Table 4 presents estimates of ϕ based on the Dunson and Baird (2001), current status and two-step approaches. Note that we have a value of C for all individuals, whether or not T_2 is observed, allowing us to implement the current status method. In the two-step and current status approaches, proportional hazards models were assumed for $S_1(t|Z)$ and $S_2(t|Z)$ as in (3) and (6), respectively. In the Dunson and Baird (2001) approach, this assumption was made for $S_1(t|Z)$ and $Q^*(t|Z)$ as in (3) and (5), respectively; S_{01} and Q_0^* were estimated as described in §3. The confidence interval for the two-step estimator is based on 500 simple bootstrap replicates; for the current status approach the analogous interval is obtained from known asymptotic theory for a coefficient estimator in a proportional hazards model based on current status data (Huang, 1996); finally, for the Dunson and Baird (2001) estimator, the interval is an (asymptotic) likelihood ratio interval, based on the profile likelihood for ϕ. TABLE 4 ABOUT HERE

The three estimates of ϕ are qualitatively similar, all reflecting little evidence of a change in risk for fibroids associated with increases in serum TCDD levels. The Dunson and Baird estimator is the largest in magnitude, whereas the two-step estimator has the smallest estimate of variability.

Figure 1 displays an estimate of $Q(t|z=0)$ obtained from the SWHS data. This estimate was obtained from estimates of the baseline survival functions S_{02} and S_{01} using the two-step method, which, in turn, provide estimates of F_{02} and F_{01} and, thus, their ratio. The plot in Figure 1 displays the estimated ratio $\hat{Q}(t|z=0)$ at times given by the values of C observed in the data. Analogous plots at $z \in \{z_{0.25}, z_{0.50}, z_{0.75}\}$ were nearly identical to that at baseline ($z=0$) and thus are not displayed. The plot in Figure 1 suggests the assumption of monotonicity of Q is violated in the SWHS data. Given the preliminary interpretation of the simulation results in §3, this lack of monotonicity suggests that the Dunson and Baird estimate in Table 1 may suffer from bias, and supports the observed precision increase enjoyed by the two-step estimator in comparison with the simpler current status approach. FIGURE 1 ABOUT HERE

As can be seen in Figure 1, at some values of t, the estimate $\hat{Q}(t|z=0)$ slightly exceeds 1 (approximately from age 25 to 32), reflecting that the estimated marginal distribution functions $\hat{F}_1(t) > \hat{F}_2$ over this age range, violating the stochastic ordering of T_1 and T_2. It is clearer from Figure 2 just how minor this ordering violation is. This shows that estimates of $F_1(t)$ are just barely less than those of $F_2(t)$ for t between 25 and 32 years. If interest focuses entirely on estimated regression coefficients (and this is measured far more precisely than the underlying distribution functions), then this minor violation in ordering is not a major issue. FIGURE 2 ABOUT HERE

5 Discussion

Our findings indicate that, in addition to its advantage of being computationally much simpler, the two-step estimator is a better choice than the approach suggested by Dunson and Baird (2001) when the monotonicity of $Q(t|Z)$ is in doubt; further the two-step approach simultaneously provides estimates of the marginal distribution functions for T_1 and T_2 that can be directly examined to assess the shape of $Q(t|Z)$. A potential disadvantage of the two-step estimator is that it does not always yield estimated onset and diagnosis distribution functions that satisfy the assumed stochastic ordering for all values of t, Z. However, the ordering violation in the application to the SWHS data is slight and occurs in distribution function estimators that are converging very slowly; thus, this issue has very little practical implication, particularly when interest is focused on estimation of regression coefficients. We note again that a complete asymptotic theory for the two-step estimator, with an appropriate estimator of asymptotic variance, remains to be established.

Based on our simulation study, the estimate of ϕ based on the two-step approach only moderately outperforms that of the simple current status estimator. Considering the simplicity of estimation of the latter, and its associated well-understood asymptotic behavior, the simple current status approach remains an attractive technique in comparison to more complicated alternatives when C is observed for all subjects. However, in many applications,

C is only observed for individuals for whom T_2 is censored; in such cases, the simple current status estimator cannot be applied and the two-step procedure is generally to be preferred over the Dunson and Baird (2001) estimator, as discussed above.

Finally, as mentioned in §4, the way missing current status data is handled across the three approaches differs slightly, and comparisons in performance may subsequently be affected. As noted by Dunson and Baird (2001), methods that incorporate data on T_2 (e.g. their approach and the two-step) in such instances would, in theory, have an advantage over the simple current status approach in situations where current status missingness depends on diagnostic history.

Supplementary Materials R code for implementing the two-step method is available under the Paper Information link at the Biometrics website http://www.tibs.org/biometrics.

Acknowledgments The authors wish to thank S. C. Shiboski for helpful conversations on computational issues and B. Eskenazi, M. Warner, P. Mocarelli, L. Needham and D. Patterson for access to data from the Seveso Women's Health Study. The second author appreciates the support of the Rockefeller Foundation through appointment as a Fellow at the Bellagio Study Center during the final stages of the writing of this paper.

References

Andrews, C., van der Laan, M. and Robins, J. (2005). Locally efficient estimation of regression parameters using current status data. *Journal of Multivariate Analysis* **96**, 332-351.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972) *Statistical Inference Under Order Restrictions.* New York: Wiley.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B* **34**, 187-220.

Dinse, G.E. and Lagakos, S.W. (1982). Nonparametric estimation of the lifetime and disease onset distributions from incomplete observations. *Biometrics* **38**, 921-32.

Dunson, D.B. and Baird, D.D. (2001). A flexible parametric model for combining current status and age at first diagnosis data. *Biometrics* **57**, 396-403.

Eskenazi, B., Mocarelli, P., Warner, M., Samuels, S., Vercellini, P., Olive, D., Needham, L., Patterson, D. and Brambilla, P. (2000). Seveso Women's Health Study: a study of the effects of 2,3,7,8-tetrachlorodibenzo-*p*-dioxin on reproductive health. *Chemosphere* **40**, 1247-53.

Eskenazi, B., Warner, M., Samuels, S., Young, J., Gerthoux, P., Needham, L., Patterson, D., Castorina, R., Olive, D., Gavoni, N., Vercellini, P., and Mocarelli, P. (2007). Serum dioxin concentrations and risk of uterine fibroids in the Seveso Women's Health Study. *American Journal of Epidemiology* in press.

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

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics* **24**, 540-68.

Jewell, N.P. and van der Laan, M. (2004). Current status data: Review, recent developments and open problems. In *Advances in Survival Analysis* , Handbook in Statistics #23, 625-42, Amsterdam: Elsevier.

Jewell, N.P., van der Laan, M. and Lei, X. (2005). Bivariate current status data with univariate monitoring times. *Biometrika* **92**, 847-862.

Kodell, R.L., Shaw, G.W. and Johnson, A.M. (1982). Nonparametric joint estimators for disease resistance and survival functions in survival/sacrifice experiments. *Biometrics* **38**, 43-58.

Lee, B.L., Kosorok, M.R. and Fine, J.P. (2005). The profile sampler. *Journal of the American Statistical Association* **100**, 960-969.

Ma, S. and Kosorok, M.R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis* **96**, 190-217.

Ma, S. and Kosorok, M.R. (2006). Adaptive penalized M-estimation with current status data. *Annals of the Institute of Statistical Mathematics* **58**, 511-526.

Shiboski, S.C. (1998). Generalized additive models for current status data. *Lifetime Data Analysis* **4**, 29-50.

Strawderman, R.L. (2006). A regression model for gap times. *The International Journal of Biostatistics* **2**, http://www.bepress.com/ijb/vol2/iss1/1.

Turnbull, B.W. and Mitchell, T.J. (1984). Nonparametric estimation of the distribution of time to onset for specific diseases in survival/sacrifice experiments. *Biometrics* **40**, 41-50.

van der Laan, M., Jewell, N.P. and Petersen, D. (1997). Efficient estimation of the lifetime and disease onset distribution. *Biometrika* **84**, 539-54.

van der Laan, M. and Jewell, N.P. (2003). Current status data and right-censored data structures when observing a marker at the censoring time. *Annals of Statistics* **31**, 512-35.

van der Laan, M.J. and Robins, J.M. (2003) *Unified Methods for Censored Longitudinal Data and Causality* New York: Springer.

Table 1: Properties of two-step, current status (CS), and Dunson and Baird (DB) estimators of ϕ and relative efficiency (RE) with two-step always in the numerator, based on ratios of mean squared error (MSE), for $n = 500$ based on 500 replicates in each of four simulations.

sim	ϕ	Q	method	E($\hat{\phi}$)	var($\hat{\phi}$)	MSE($\hat{\phi}$)	RE
1	-0.4	monotonic	2-step	-0.4285	0.0195	0.0203	1
			CS	-0.4300	0.0193	0.0202	1.006
			DB	-0.4012	0.0164	0.0164	1.240
2	0.0	monotonic	2-step	-0.0016	0.0196	0.0197	1
			CS	-0.0020	0.0202	0.0202	0.972
			DB	-0.0030	0.0174	0.0174	1.129
3	-0.4	non-monotonic	2-step	-0.4038	0.0302	0.0302	1
			CS	-0.4219	0.0327	0.0332	0.909
			DB	-0.5175	0.0265	0.0403	0.750
4	0.0	non-monotonic	2-step	-0.0115	0.0243	0.0244	1
			CS	-0.0074	0.0285	0.0285	0.856
			DB	-0.2244	0.0232	0.0735	0.332

Table 2: Variance of the two-step estimator of ϕ, median of the simple bootstrapped estimates of this variance (varbs) and weighted bootstrapped estimates of this variance (varbsw) based on 500 replicates in each of four simulations.

sim	$\mathrm{var}(\hat{\phi})$	$\mathrm{median}\{\mathrm{varbs}(\hat{\phi})\}$	$\mathrm{median}\{\mathrm{varbsw}(\hat{\phi})\}$
1	0.0195	0.0199	0.0181
2	0.0196	0.0205	0.0189
3	0.0302	0.0294	0.0289
4	0.0244	0.0233	0.0226

Table 3: Properties of two-step estimator of ϕ when true survival function for T_2 is misspecified. True form is $S_2(t|Z,W) = S_{02}^{\exp(\beta Z+\gamma W)}$, where $\gamma = -3.0$. Estimation erroneously excludes W in estimation of S_2.

sim	ϕ	E($\hat{\phi}$)	var($\hat{\phi}$)	MSE($\hat{\phi}$)
1	-0.4	-0.4195	0.0211	0.0215
2	0.0	0.0062	0.0200	0.0201
3	-0.4	-0.4056	0.0334	0.0334
4	0.0	0.0130	0.0290	0.0292

Table 4: Estimate of ϕ and 95% confidence interval based on the two-step, current status (CS) and Dunson and Baird (DB) approaches.

Method	$\hat{\phi}$	95% confidence interval
2-step	-0.07	$(-0.26, 0.10)$
CS	-0.11	$(-0.36, 0.13)$
DB	-0.18	$(-0.43, 0.07)$

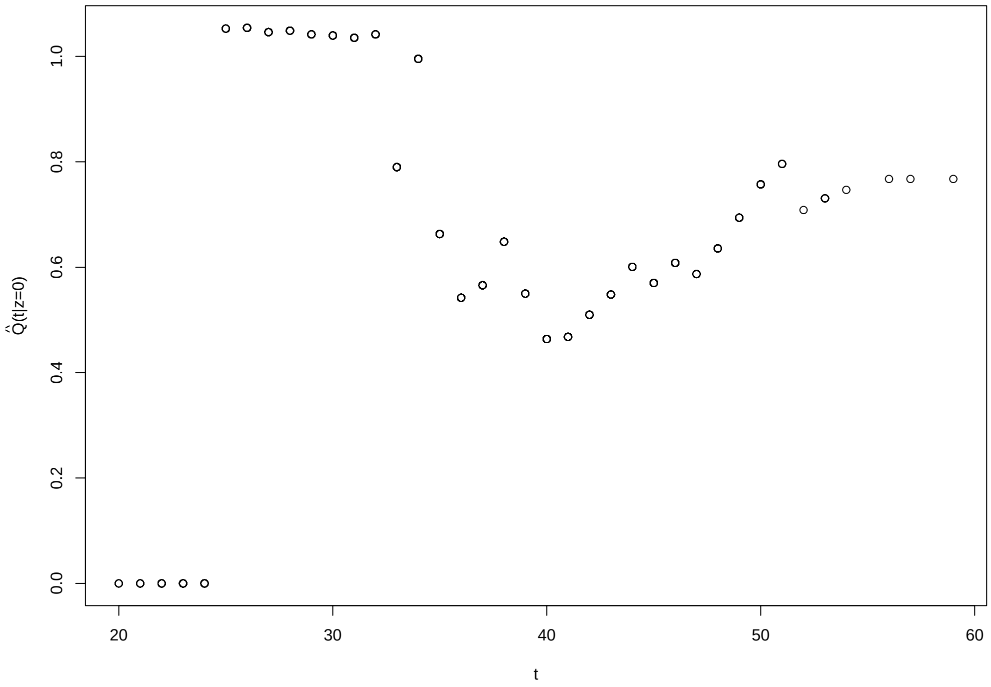

Figure 1: $\hat{Q}(t|z=0)$ estimated using two-step approach for SWHS data; $z = \log_{10}(\text{TCDD})$. The time scale t is age in years.

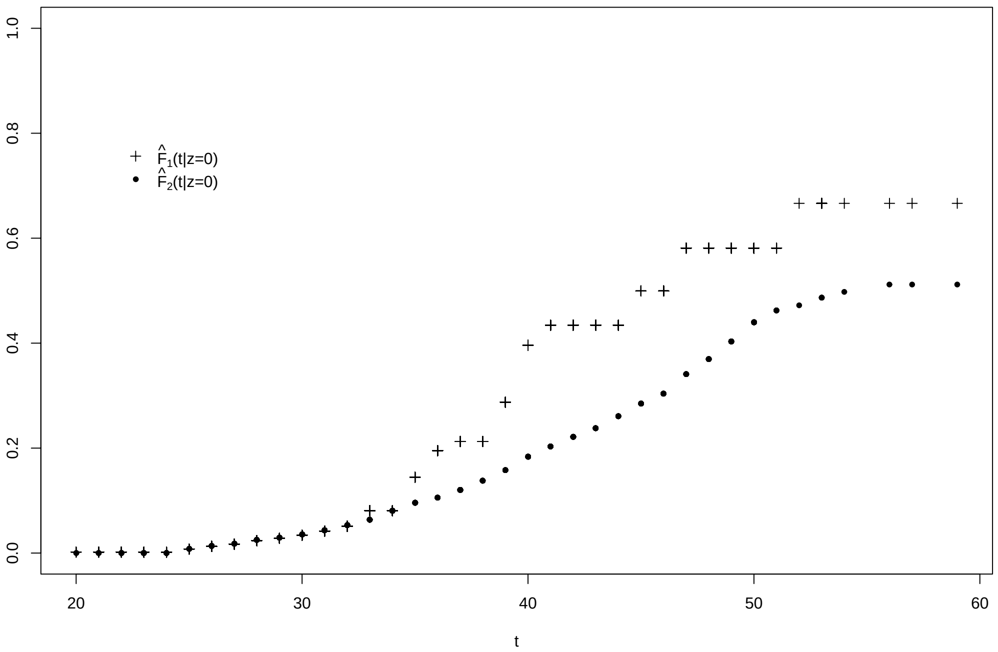

Figure 2: $\hat{F}_1(t|z = 0)$ and $\hat{F}_2(t|z = 0)$ estimated using two-step approach for SWHS data; $z = \log_{10}(\text{TCDD})$. The time scale t is age in years.