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

Year 2005 *Paper* 195

Correspondences between Regression Models for Complex Binary Outcomes and Those for Structured Multivariate Survival Analyses

Nicholas P. Jewell*

*Division of Biostatistics, School of Public Health, University of California, Berkeley, jewell@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper195

Correspondences between Regression Models for Complex Binary Outcomes and Those for Structured Multivariate Survival Analyses

Nicholas P. Jewell

Abstract

Doksum and Gasko [5] described a one-to-one correspondence between regression models for binary outcomes and those for continuous time survival analyses. This correspondence has been exploited heavily in the analysis of current status data (Jewell and van der Laan [11], Shiboski [18]). Here, we explore similar correspondences for complex survival models and categorical regression models for polytomous data. We include discussion of competing risks and progressive multi-state survival random variables.

1 Introduction

Consider a continuous time survival response T that is measured on an individual with a known p-dimensional set of covariates $\mathbf{Z} = (Z_1, \ldots, Z_p)$. A survival regression model focuses on the relationship between the (conditional) distribution function $F_{\mathbf{z}}$, of T, given $\mathbf{Z} = \mathbf{z}$, and $\mathbf{z}$. Examples of such models include the Cox model (Cox [3]) and the proportional odds model (Bennett [2]); in each of these, the model can be made fully parametric if both the regression relationship and a baseline version of F are parametrically described, semi-parametric if only one of these is parametrically modeled, or nonparametric if both are only loosely specified.

At a fixed value t, a binary characteristic is defined by the event $T < t$ which occurs with probability $p_t = \Pr(T < t)$. A survival time regression model for T automatically induces a binary regression model relating $p_{t;\mathbf{z}} = \Pr(T < t|\mathbf{Z} = \mathbf{z})$ to both t and $\mathbf{z}$. Doksum and Gasko [6] examine this correspondence in detail for several familiar survival models including those noted above. One simple example is the proportional odds model for which

$$F_{\mathbf{z}}(t) = \frac{e^{\alpha(t)+\beta\mathbf{z}}}{1 + e^{\alpha(t)+\beta\mathbf{z}}} \tag{1}$$

where $\alpha(t)$ is a non-decreasing function with $\alpha(0) = -\infty, \alpha(\infty) = \infty$, and where β is a p-dimensional vector of regression coefficients. With this model

$$\log \frac{p_{t;\mathbf{z}}}{1 - p_{t;\mathbf{z}}} = \alpha(t) + \beta\mathbf{z},$$

a logistic regression model in $\mathbf{z}$ with 'intercept' $\alpha(t)$. As noted, if $F_{\mathbf{0}}$ is assumed to follow a parametric model (for example, the log-logistic distribution [15], Chapter 2.2.6), then this logistic model is also fully parametric (with $\alpha(t) = a + b \log t$ in the log-logistic distribution case with a and $b > 0$ suitably chosen).

This correspondence between survival time and binary outcome regression models has been heavily exploited in the analysis of current status data. Current status observation refers to a form of incompleteness in that the data consists of independent observations on the random variable $(C, \Delta = I(T < C), \mathbf{Z})$ instead of $(T, \mathbf{Z})$. Here, C can be random or deterministic, and is usually referred to as the *monitoring time*; it is typically assumed that C is independent of T and is uninformative. The variable Δ indicates the *current status* of an individual at time C, namely whether $T < C$ or not. A review of various forms and examples of current status data is given in Jewell and van der Laan [14] and the references contained therein.

As follows from the work of Doksum and Gasko [6], a regression model for the (unobserved) T immediately leads to a binary regression model for the observed binary outcome

Δ with C included as an additional covariate along with $\mathbf{Z}$. For example, with the proportional odds model for T in (1.1), the model for Δ is given by the logistic regression

$$\log \frac{p_{c;\mathbf{z}}}{1-p_{c;\mathbf{z}}} = \alpha(c) + \beta\mathbf{z}, \tag{2}$$

where $p_{c;\mathbf{z}} = \Pr(\Delta = 1|C = c, \mathbf{Z} = \mathbf{z})$, and $\alpha(c)$ is necessarily non-decreasing in c with $\alpha(0) = -\infty$ and $\alpha(\infty) = \infty$.

There are two primary properties of this correspondence of regression models for current status data that make it particularly useful. First, for many standard survival models, the effects of C and $\mathbf{Z}$ are additive in the binary regression setting when the appropriate link function is used. This is illustrated by the proportional odds model when a logistic link is used for Δ as shown in (1.2). A similar property holds for the proportional hazards model with the complimentary log-log link for Δ (see Jewell and van der Laan [14]). Second, the parameter β in the binary regression model for the observed Δ has an immediate interpretation in the survival regression model for the unobserved T. For example, in the logistic regression model (1.2), the regression coefficient β_k is nothing more than the log odds ratio of failure by time t, associated with a unit increase in the k^{th} component of $\mathbf{Z}$ (holding other component variables constant), as given by the original proportional odds model (1.1). Although this assumes no interaction terms in $\mathbf{Z}$, the ideas and interpretations immediately generalize to more complex models. For more detail on the application of the proportional odds model in the context of current status data see Rossini and Tsiatis [20].

The purpose of this article is to examine analogous correspondences between regression models for more complex survival data and their current status counterparts. We pay specific attention to two settings: (i) competing risks survival models which naturally lead to unordered polytomous current status outcomes (Section 2), and (ii) progressive multi-state survival models, corresponding to ordinal categorical current status outcomes (Section 3). While such correspondences naturally extend to these more complex settings, we show that the attractive features of additivity in C and $\mathbf{Z}$, and interpretability of regression coefficients, only can be guaranteed with additional assumptions, at least in the models considered here. In Section 4, we therefore briefly discuss the advantages (and disadvantages) of modeling marginal distributions separately using the simpler survival and binary regression connections for standard current status data.

2 Competing Risks Survival Models and Polytomous Regression Models

Competing risks survival data arise in situations where, in addition to observations on the failure time T, there is also information on a categorical variable J which takes on the values $1, \ldots, m$, and represents the cause or type of failure at time T. It is standard to assume that all failures are associated with one and only one value of J. The joint distribution of the random variable (T, J) is of primary interest. See Crowder [4] for a recent treatment of the topic.

The cause-specific hazard function for cause $J = 1, \ldots, m$, (Kalbfleisch and Prentice [15]) is defined by

$$\lambda_j(t) = \lim_{h \to 0} h^{-1} \Pr[t \leq T < t + h, J = j | T \geq t].$$

Related to these cause-specific hazards are the sub-distribution functions of primary interest given by

$$F_j(t) = \Pr(T < t, J = j), \qquad j = 1, \ldots, m$$

with the overall survival function then

$$S(t) = 1 - \sum_{j=1}^{m} F_j(t).$$

Note that the cause-specific density function

$$f_j(t) = \lim_{h \to 0} h^{-1} \Pr[t \leq T < t + h, J = j],$$

is the derivative of F_j. Finally, these functions are related through

$$f_j(t) = \lambda_j(t) F(t)$$

where $F(t) = 1 - S(t) = F_1(t) + \cdots + F_m(t)$.

We now introduce the covariate $\mathbf{Z}$ into the notation, writing for example

$$F_j(t; \mathbf{z}) = \Pr(T < t, J = j | \mathbf{Z} = \mathbf{z}),$$

for $j = 1, \ldots, m$, with

$$F_0(t; \mathbf{z}) = \Pr(T \geq t | \mathbf{Z} = \mathbf{z}) = \mathbf{S}(\mathbf{t}|\mathbf{z}).$$

Before further discussion of regression models, it will be helpful to introduce an alternative description of the joint distribution of (T, J). For each j, let α_j be a non-decreasing

function on $[0, \infty)$ for which $\alpha_j(0) = -\infty$ and $\alpha_j(\infty) = \infty$. Further, assume that these m functions are commensurate in the sense that the functions

$$\frac{e^{\alpha_j(t)}}{1 + \sum_{k=1}^{m} e^{\alpha_k(t)}} \tag{3}$$

are non-decreasing, for $j = 1, \ldots, m$. Then

$$F_j(t) = \frac{e^{\alpha_j(t)}}{1 + \sum_{k=1}^{m} e^{\alpha_k(t)}} \tag{4}$$

define sub-distribution functions for m competing risks. Note that solving (2.2) yields the inverse relationships

$$\alpha_j(t) = \log\left[\frac{F_j(t)}{S(t)}\right]. \tag{5}$$

for $j = 1, \ldots, m$. We can thus characterize the joint distribution of (T, J) equally well in terms of either $\{\alpha_1, \ldots, \alpha_m\}$ or $\{F_1, \ldots, F_m\}$, with the appropriate constraints on either set of functions.

We are now in a position to describe a natural regression model for $F_1, \ldots, F_m$. For each $j = 1, \ldots, m$ and each covariate value $\mathbf{z}$ we write

$$F_j(t; \mathbf{z}) = \frac{e^{\alpha_j(t) + \beta_j \mathbf{z}}}{1 + \sum_{k=1}^{m} e^{\alpha_k(t) + \beta_k \mathbf{z}}}, \tag{6}$$

where β_j is a $1 \times p$ vector of regression coefficients. This model introduces the key additive separation of the effects of t and z on the sub-distribution functions that we noted was valuable in the standard setting. We refer to this model as the *proportional odds model with competing risks* as it generalizes the model of the same name in the single risk setting (Bennett [2]). Before proceeding further, however, for (2.4) to describe a set of sub-distribution functions, we need the functions $\{\alpha_j^* = \alpha_j(t) + \beta_j \mathbf{z} : j = 1, \ldots, m\}$ to satisfy the constraints, described by (2.1), assuming that $\{\alpha_j : j = 1, \ldots, m\}$ do. Trivially $\{\alpha_j^*; j = 1, \ldots, m\}$ possess the same limits as $\{\alpha_j; j = 1, \ldots, \alpha_m\}$ at both 0 and ∞. In considering the constraints (2.1), we consider the case where $m = 2$ for simplicity.

Differentiating (2.1) with respect to t shows that (2.1) is equivalent to

$$\begin{aligned} \alpha_1'(t) + e^{\alpha_2(t)}[\alpha_1'(t) - \alpha_2'(t)] &\geq 0, \\ \alpha_2'(t) + e^{\alpha_1(t)}[\alpha_2'(t) - \alpha_1'(t)] &\geq 0, \end{aligned}$$

for all t. Therefore, for (2.4) to correspond to a survival model for competing risks for any value of β and $\mathbf{z}$, we need

$$\begin{aligned} \alpha_1'(t) + e^{a_2} e^{\alpha_2(t)}[\alpha_1'(t) - \alpha_2'(t)] &\geq 0, \\ \alpha_2'(t) + e^{a_1} e^{\alpha_1(t)}[\alpha_2'(t) - \alpha_1'(t)] &\geq 0, \end{aligned}$$

for all t and any value of $a_1 = \beta_1 \mathbf{z}$ and $a_2 = \beta_2 \mathbf{z}$. Without further restrictions on α_1 and α_2, this holds if and only if $\alpha_1'(t) - \alpha_2'(t) = 0$ for all t. Noting that $\alpha_1'(t) = [F_1 S]^{-1}[f_1 - f_1 F_2 + f_2 F_1]$ (where $F_1(t) = F_1(t; \mathbf{0})$, etc), with an analogous expression for $\alpha_2'(t)$, it follows that

$$\alpha_1'(t) - \alpha_2'(t) = \frac{f_1}{F_1} - \frac{f_2}{F_2} = \left(\log \frac{F_1}{F_2} \right)'.$$

Thus $\alpha_1'(t) - \alpha_2'(t) = 0$ is equivalent to F_1 and F_2 being proportional, in turn, equivalent to proportionality of the two cause-specific hazard functions λ_1 and λ_2.

In sum, we have shown that the proportional odds regression model (2.4) only yields proper sub-distribution functions F_j for all values of β and $\mathbf{z}$ if the cause-specific hazards are proportional for all values of $\mathbf{z}$, a very restrictive condition. With this assumption, however, the parameters α_j and β_j have specific interpretations as follows. First, it follows from (2.3) that, for individuals at the baseline level of $\mathbf{Z} = \mathbf{0}$, $\alpha_j(t)$ is just the log odds, at time t, that a failure of type j has occurred as against no failure. Further, from (2.4) it follows that

$$\frac{F_j(t; \mathbf{z})}{S(t; \mathbf{z})} = e^{\alpha_j(t) + \beta_j \mathbf{z}},$$

so that the k^{th} component of the regression coefficient β_j is the log odds of failure by time t, due to cause j, as against no failure, associated with a unit increase in the k^{th} component of $\mathbf{Z}$ (holding other component variables constant). This is the case at all values of t. Similarly, note that

$$\frac{F_j(t; \mathbf{z})}{F_k(t; \mathbf{z})} = e^{\alpha_j(t) - \alpha_k(t)} e^{(\beta_j - \beta_k)\mathbf{z}},$$

showing that the log odds of failure by time t due to cause j, as against failure by time t due to cause k, is linear in $\mathbf{z}$ with slope $\beta_j - \beta_k$, again true for all t.

Given the restriction of proportional cause-specific hazards, why is the model (2.4) appealing in the first place? The answer is in its relationship to a regression model for a polytomous outcome generated by a current status observation scheme. Specifically, suppose that, for each individual, information on survival status, and, if relevant, cause of failure, is available only at a single time C. Thus, the observed data can be represented as (C, Δ), where $\Delta = 0$ if $T \geq C$, $\Delta = j$ if $T < C$ with $J = j$, for $1 \leq j \leq m$. It is therefore assumed that if an individual is known to have failed at the observation time C, the cause of failure is also available. As before, we assume that the monitoring time C is independent of T and is uninformative.

Note first that the distribution of Δ is related to that of (T, J) simply as follows:

$$\Pr(\Delta = j | C) = F_j(C), \tag{7}$$

for $j = 1, \ldots, m$ with $\Pr(\Delta = 0|C) = S(C)$.

For a fixed C, it is natural to consider a regression model which links the distribution of Δ to covariates $\mathbf{Z}$. A natural model is the *multinomial logistic model* which describes the dependence of $\Pr(\Delta = j|C, \mathbf{Z} = \mathbf{z})$ on the explanatory variables. In particular, the model states that

$$\Pr(\Delta = j|C, \mathbf{Z} = \mathbf{z}) = \frac{e^{\alpha_j + \beta_j \mathbf{z}}}{1 + \sum_{k=1}^{m} e^{\alpha_k + \beta_k \mathbf{z}}}, \quad j = 1, \ldots, m, \tag{8}$$

with necessarily

$$\Pr(\Delta = 0|C, \mathbf{Z} = \mathbf{z}) = \frac{1}{1 + \sum_{k=1}^{m} e^{\alpha_k + \beta_k \mathbf{z}}}.$$

See, for example, McCullagh and Nelder [18], Chapter 5.2.4.

Extending this model to allow for varying C, while ensuring additivity of effects of C and $\mathbf{Z}$, immediately suggests replacing α_j with $\alpha_j(C)$ in (2.6), where the functions α_j satisfy the constraints given in (2.1) along with appropriate limits. Through (2.5) and (2.6), this immediately corresponds to the proportional odds model (2.4) for (T, J). As a consequence of our analysis of (2.4), this shows that we can only 'properly' use the multinomial logistic model for current status competing risks data, with additive effects of C and the covariates, if we are willing to assume that the underlying cause-specific hazards are proportional. Even in this restrictive case, it is important to note that practical issues remain for joint estimation of α and $\beta_1, \ldots, \beta_m$, particularly when α is treated nonparametrically.

2.1 The Proportional Hazards Model

Extending the ubiquitous Cox proportional hazards model (Cox [3]), the proportional hazards model for competing risks (Crowder [4], Chapter 1.4.1; Kalbfleisch and Prentice [15], Chapter 8.12) specifies that the conditional cause-specific hazard functions satisfy

$$\lambda_j(t; \mathbf{z}) = \lambda_{0j}(t) e^{\beta_j \mathbf{z}}, \tag{9}$$

for $j = 1, \ldots, m$, where λ_{0j} is the baseline cause-specific hazard function for cause j for individuals with $\mathbf{z} = \mathbf{0}$. This should not be confused with the assumption of proportional cause-specific hazards, at any fixed value of $\mathbf{Z}$, that we discussed earlier in Section 2, and that we return to briefly below.

It is of interest to determine the form of polytomous regression model that the proportional hazards model for (T, J), given in (2.7), induces on current status observations

(C, Δ). First, without covariates, note that

$$\Pr(\Delta = j|C) = \int_0^C \lambda_j(u) \exp\left[-\int_0^u \left(\sum_{k=1}^m \lambda_k(t)\right) dt\right] du.$$

Now introducing the covariates $\mathbf{Z}$, under (2.7), we have

$$\Pr(\Delta = j|C, \mathbf{Z} = \mathbf{z}) = \int_0^C \lambda_{0j}(u) e^{\beta_j \mathbf{z}} \prod_{k=1}^m \exp\left[-\int_0^u \lambda_{0k}(u) e^{\beta_k \mathbf{z}} dt\right] du. \tag{10}$$

This explicitly links the proportional hazards model for (T, J) to a multinomial regression model for the current status observation (C, Δ), albeit a rather cumbersome one. In particular, there appears to be no convenient link function which separates the right hand side of (2.8) into additive effects for C and $\mathbf{z}$. It is plausible that further assumptions might lead to a simpler relation than (2.8). Suppose, for example, we now additionally assume proportional cause-specific hazard functions, so that, in particular, $\lambda_{0j}(t) = a_j \lambda_0(t)$ for all t and $j = 1, \ldots, m$, where the a_j's are positive constants and λ_0 is an unspecified hazard function. Then (2.8) simplifies to

$$\begin{aligned}
\Pr(\Delta = j|C, \mathbf{Z} = \mathbf{z}) &= a_j e^{\beta_j \mathbf{z}} \int_0^C \lambda_0(u) \prod_{k=1}^m \exp\left[e^{-a_k e^{\beta_k \mathbf{z}}} \int_0^u \lambda_0(t) dt\right] du \\
&= \frac{a_j e^{\beta_j \mathbf{z}}}{\sum_{k=1}^m a_k e^{\beta_k \mathbf{z}}} \times \\
&\quad \left[1 - \exp\left\{\left(-\sum_{k=1}^m a_k e^{\beta_k \mathbf{z}}\right) \int_0^C \lambda_0(t) dt\right\}\right]. \qquad (11)
\end{aligned}$$

Note that, for simplicity, we can absorb the constants $a_1, \ldots, a_m$ into the regression terms so long as a constant is included in $\mathbf{Z}$, yielding

$$\Pr(\Delta = j|C, \mathbf{Z} = \mathbf{z}) = \frac{e^{\beta_j \mathbf{z}}}{\sum_{k=1}^m e^{\beta_k \mathbf{z}}} \left[1 - \exp\left\{\left(-\sum_{k=1}^m e^{\beta_k \mathbf{z}}\right) \int_0^C \lambda_0(t) dt\right\}\right],$$

where we adjust our definition and interpretation of $\beta_1, \ldots, \beta_m$. However, the main point is that the effects of C and $\mathbf{z}$ remain inextricably linked in (2.9), even when further restrictions are placed on the shape of λ_0. The closest analogue, arising from (2.9), to the univariate correspondence of the proportional hazards model to a complementary log-log regression model for Δ, is that

$$\begin{aligned}
\log\left[-\log\left\{\Pr(T > C|J = j, C, \mathbf{z})\right\}\right] &= \log\left[-\log\left\{1 - \frac{F_j(C; \mathbf{z})}{F_j(\infty; \mathbf{z})}\right\}\right] \\
&= \log\left(\sum_{k=1}^m e^{\beta_k \mathbf{z}}\right) + \log \Lambda_0(C), \qquad (12)
\end{aligned}$$

where Λ_0 is the integrated hazard function associated with λ_0. Unfortunately, $\Pr(T > C|J = j, C, \mathbf{z})$, in the left hand side of (2.10), does not obviously correspond to any (conditional) expectation of an observable random variable with current status data (except where the cause of failure is also observed for those for whom the failure event has not occurred at time C). Even then, the right hand side, while showing additivity of the effects for C and $\mathbf{z}$, does not yield a simple linear term in $\mathbf{z}$ when $m > 1$.

In sum, although the proportional hazards model for competing risks data necessarily induces a multinomial regression model for the categorical data produced by current status observation, the resulting model does not simply correspond to a recognizable multinomial regression model which might allow the use of existing software (possibly adapted to allow for monotonicity constraints in the nonparametric case). Similarly, application of a 'standard' generalized linear model for nominal multinomial outcomes to current status observations of competing risks data cannot be simply interpreted in terms of an underlying proportional hazards model even with additional restrictive assumptions.

2.2 Mixture Models for Competing Risks

Larson and Dinse [17] suggested a mixture model for competing risks data which, in its simplest form, is as follows. First, a multinomial logistic regression model is assumed for $F_j(\infty; \mathbf{z})$, the fraction of all eventual events from cause j, so that

$$F_j(\infty; \mathbf{z}) = \frac{e^{\alpha_j \mathbf{z}}}{\sum_{k=1}^{m} e^{\alpha_k \mathbf{z}}},$$

for some set of regression coefficients $\alpha_1, \ldots, \alpha_m$, where a constant term is included in $\mathbf{Z}$, and for identifiability we assume, for example, $\alpha_1 = 0$. The second part of the model specifies regression relationships for the conditional distribution functions $H(T|J)$ that determine properties of event times associated with each specific cause. In particular, a proportional hazards model for these distribution functions yields $1 - H(t|J = j; \mathbf{Z} = \mathbf{z}) = \exp\left(\Lambda_j(t) e^{\beta_j \mathbf{z}}\right)$ for some set of integrated hazard functions $\Lambda_j, j = 1, \ldots, m$, so that

$$\Pr(\Delta = j|C, \mathbf{Z} = \mathbf{z}) = \frac{e^{\alpha_j \mathbf{z}}}{\sum_{k=1}^{m} e^{\alpha_k \mathbf{z}}} \left[1 - \exp\left(-e^{\beta_j \mathbf{z}} \int_0^C \lambda_0(t) dt\right)\right]. \tag{13}$$

Note the similarity with (2.9). Again we can rewrite (2.11) to obtain the analogue of (2.10), namely

$$\log\left[-\log\left\{\Pr(T > C|J = j, C, \mathbf{z})\right\}\right] = \log\left[-\log\left\{1 - \frac{F_j(C; \mathbf{z})}{F_j(\infty; \mathbf{z})}\right\}\right] = \beta_j \mathbf{z} + \log \Lambda_0(C).$$

This yields additive effects for C and $\mathbf{z}$, and now a linear term in $\mathbf{z}$ on the right hand side, but, of course, suffers from the same drawback as (2.10) in that the left hand side does not correspond to the (conditional) expectation of an observable random variable with current status data.

3 Progressive Multi-State Survival Models and Ordinal Polytomous Regression Models

We now turn to generalizations of a simple survival random variable in a quite different direction. Suppose interest focuses on a finite state survival process where individuals have to successively progress through each of $m+1$ states over time. The illness-death model is a special case of this scenario with $m = 2$. Specifically, let $X(t)$ be a counting process with m jump times denoted by the random variables $T_1, \ldots, T_m$, where necessarily $T_1 \leq T_2 \leq \cdots \leq T_m$. We wish to understand the joint distribution, F, of $(T_1, \ldots, T_m)$ and the influence of explanatory variables on its properties. We focus here solely on models for the marginal distributions of F, denoted by $F_1, \ldots, F_m$ since only these marginals are identifiable from current status data. One immediate consequence of this is that the constraint $\Pr(T_1 \leq \cdots \leq T_m) = 1$ does not imply a stronger constraint on the marginals other than that $F_1 \geq \cdots \geq F_m$. This follows, since for any set of marginal distributions $F_1, \ldots, F_m$ with $F_1 \geq \cdots \geq F_m$, there exists an m-dimensional distribution with $\Pr(T_1 \leq \cdots \leq T_m) = 1$ that has marginals $F_1, \ldots, F_m$. To keep things simple, we also assume throughout that $F_1, \ldots, F_m$ are all continuous.

As in Section 2 we first consider the scenario absent covariates, and introduce a useful parameterization of F. For $j = 1$, let $\alpha_1(t)$ be defined by

$$\alpha_1(t) = \log\left[\frac{1 - F_1(t)}{F_1(t)}\right]. \tag{14}$$

Now consider the conditional probabilities of T_j, given T_{j-1}, for $j > 1$. In particular, define

$$\alpha_j(t) = \log\left[\frac{G_j(t)}{(1 - G_j(t))}\right], \tag{15}$$

where

$$G_j(t) = \Pr(T_j \geq t | T_{j-1} < t) = \frac{F_{j-1}(t) - F_j(t)}{F_{j-1}(t)} = 1 - \frac{F_j(t)}{F_{j-1}(t)}, \tag{16}$$

for $1 < j \leq m$. We can solve (3.1–3.3) for F_j, giving

$$F_j = \prod_{k=1}^{j} \frac{1}{1 + e^{\alpha_k}}. \tag{17}$$

The functions α_j re-express the marginal distributions $F_1, \ldots, F_m$, and necessarily have to satisfy appropriate conditions for (3.4) to yield proper distribution functions. The conditions for α_1 are straightforward in that $\alpha_1(0) = \infty$, $\alpha_1(\infty) = -\infty$ and α_1 is non-increasing. For α_j with $j > 1$, the constraints are more complex. Formally, $\alpha_j(\infty) = -\infty$, and the α_js possess the mutual properties that the functions $\prod_{k=1}^{j} \frac{1}{1+e^{\alpha_k}}$ are all non-decreasing for $j = 1, \ldots, m$. For example, with $j = 2$, this requires that

$$\alpha_1{}' e^{\alpha_1} + \alpha_2{}' e^{\alpha_2} + (\alpha_1{}' + \alpha_2{}') e^{\alpha_1 + \alpha_2} \leq 0, \tag{18}$$

with analogous conditions for the other α_js for $j > 2$. Note that, along with the conditions on α_1, α_2 being non-increasing is a sufficient condition for (3.5); in general, α_j being non-increasing for all j implies proper distribution functions F_j (along with the appropriate limit conditions). However, it is not necessary that α_j be non-increasing. For example, with $m = 2$—the standard nonparametric illness-death model—α_2 being non-increasing is equivalent to F_2/F_1 being non-decreasing. However, suppose that, for small t, progression to illness (the first transition) is immediately followed by the second transition (to death), but for large t, there is a much longer gap between the two transitions. Then, initially F_2/F_1 is close to 1 and then decreases as t gets larger.

We now introduce regression effects of $\mathbf{Z}$ on each of $T_1, \ldots, T_m$. In principal, we cannot simply postulate separate unlinked regression models for each of $T_1, \ldots, T_m$ in turn, as this may lead to violations of the stochastic ordering of $T_1, \ldots, T_m$ for certain values of regression coefficients and/or $\mathbf{Z}$. Suppose, alternatively, that we focus on the effects of $\mathbf{Z}$ on the functions $\alpha_1, \ldots, \alpha_m$, and assume that these are linear,

$$\alpha_j(t|\mathbf{Z} = \mathbf{z}) = \alpha_j(t|\mathbf{Z} = \mathbf{0}) + \beta_j \mathbf{z}, \tag{19}$$

or equivalently,

$$F_j(t|\mathbf{Z} = \mathbf{z}) = \prod_{k=1}^{j} \frac{1}{1 + e^{\alpha_k(t|\mathbf{z})}} = \prod_{k=1}^{j} \frac{1}{1 + e^{\alpha_k(t|\mathbf{Z}=\mathbf{0}) + \beta_k \mathbf{z}}}, \tag{20}$$

for $j = 1, \ldots, m$. For (3.7) to correspond to proper distribution functions, it is necessary that the constraining conditions, exemplified by (3.5)—when $\mathbf{Z} = \mathbf{0}$—imply that the same conditions hold for $\alpha_j(t|\mathbf{Z} = \mathbf{z})$ in (3.6). However, this is not guaranteed for all values of β_j and $\mathbf{z}$ except in particular circumstances. One such is the additional assumption that

$\alpha_j(t|\mathbf{Z}=\mathbf{0})$ is non-increasing for all j, or equivalently that $F_j(t|\mathbf{Z}=\mathbf{0})/F_{j-1}(t|\mathbf{Z}=\mathbf{0})$ is non-decreasing in t for $j>1$. This additional condition implies that the regression model (3.7) always yields a set of proper distribution functions $F_j(t|\mathbf{Z})$ for all j, β_j, and any value of $\mathbf{Z}$.

We call the model (3.7) a *proportional odds model* for $T_1,\ldots,T_m$ because of the interpretation of the regression coefficient vectors β_j. Note that a unit increase in the k^{th} component of $\mathbf{Z}$ (holding other components fixed) increases the log odds of being in state j, conditional on being in state j or higher, by β_{jk}, the k^{th} component of β_j. As in the other cases we have studied, the functions $\alpha_j(t|\mathbf{Z}=\mathbf{0})$ determine the shape of the baseline distribution functions $F_j(t|\mathbf{Z})$ for $\mathbf{Z}=\mathbf{0}$, $j=1,\ldots,m$.

We now relate these ideas to current status observation on $T_1,\ldots,T_m$, at a monitoring time C. Here, the observed data can be represented as $Y=(C,\Phi)$, where $\Phi=j$ if $T_{j-1}<C\leq T_j$ for $j=1,\ldots,m+1$, where $T_0\equiv 0$ and $T_{m+1}\equiv\infty$. As before, we assume that observation times are independent of $T_1,\ldots,T_m$, and are uninformative.

For a fixed C, we again focus on models for $p_{j,\mathbf{z}}=\Pr(\Phi=j|\mathbf{Z}=\mathbf{z})$. Note that, suppressing the dependence on $\mathbf{z}$ for the moment, $p_1=\Pr(T_1\geq C)=1-F_1(C)$, $p_{m+1}=\Pr(T_m<C)=F_m(C)$, and

$$p_j(C)=\Pr(T_{j-1}<C\leq T_j)=F_{j-1}(C)-F_j(C), \tag{21}$$

for $j=2,\ldots,m$. Note that $p_{j+1}(C)+\cdots+p_{m+1}(C)=F_j(C)$ for $j=1,\ldots,m$.

A natural regression model here is the so-called *sequential logit model* for ordinal categorical data that is defined by logistic regression models for the sequential probabilities $p_{j,\mathbf{z}}/(p_{j,\mathbf{z}}+\cdots+p_{m+1,\mathbf{z}})$. In terms of log odds, this yields

$$\log\frac{p_{j,\mathbf{z}}}{p_{j+1,\mathbf{z}}+\cdots+p_{m+1,\mathbf{z}}}=\alpha_j+\beta_j\mathbf{z}, \tag{22}$$

for $j=1,\ldots,m$. This is also referred to as the *continuation ratio logit model*; see, for example, Agresti ([1], Chapter 7.4.3).

We now want to incorporate varying monitoring times C, again with the idea of assuming that the effects of C are additive to those of the covariates. This is achieved by assuming that only the intercept terms α_j depend on C, and not the slope coefficients β_j, in (3.9). The final model is therefore

$$\log\frac{p_{j,\mathbf{z}}(C)}{p_{j+1,\mathbf{z}}(C)+\cdots+p_{m+1,\mathbf{z}}(C)}=\alpha_j(C)+\beta_j\mathbf{z}. \tag{23}$$

Using (3.8), the model (3.10) therefore corresponds exactly with the proportional odds model (3.7). The consequence again is that the sequential logistic model for ordered multi-state current status data, with additive effects of C and the covariates, corresponds with the proposed proportional odds model for $T_1, \ldots, T_m$ so long as the intercept functions in C satisfy the constraints induced by the functions in (3.7) being non-decreasing, as discussed earlier, and the associated limit conditions.

The situation is therefore somewhat more satisfying than in the competing risks situation where the multinomial logistic model for current status data, with additive effects, implied that the underlying competing risks model is only proper if the intercept functions have identical derivatives (corresponding to the restrictive condition of proportional cause-specific hazards). With ordered multi-state current status data, the sequential logistic model (3.10) corresponds to any set of marginal distributions for $T_1, \ldots, T_m$, albeit with cumbersome monotonicity conditions on the intercept functions. As previously noted, the simple conditions that α_j be non-increasing for all j may be more useful in practice, but requires the additional assumption that the distribution functions $F_j(t|\mathbf{z})/F_{j-1}(t|\mathbf{z})$ are non-decreasing in t for $j = 2, \ldots, m$.

The regression model (3.7) has been previously suggested in an example concerning transitions of women from a disease-free state, to onset of pre-clinical fibroids, to diagnosis of fibroids (i.e. $m = 2$) in Dunson and Baird [7], as part of a richer data structure where T_2 is often observed directly (for the single group setting for such data, see van der Laan, Jewell and Petersen [23]). Although Dunson and Baird [7] developed the model in an ad hoc fashion, they also invoked the assumption that F_2/F_1 be non-decreasing to simplify semiparametric estimation strategies, arguing that this assumption is reasonable in the fibroid example. For previous work on current status data for multi-state stochastic processes in the single group setting, see Jewell and van der Laan [10, 11] and van der Laan and Jewell [22].

We note here that there is an obvious alternative sequential logistic model which focuses on conditional probabilities in the alternative 'direction' from (3.10). Specifically, we could sequentially use a logistic model for the probabilities $p_{j,\mathbf{z}}/(p_{1,\mathbf{z}}+\cdots+p_{j,\mathbf{z}})$ for $j = 1, \ldots, m+1$ which is linear in $\mathbf{z}$ with an additive term in C. In analogous fashion this leads to the regression model

$$S_j(t|\mathbf{Z} = \mathbf{z}) = \prod_{k=j}^{m} \frac{1}{1 + e^{\gamma_k(t|\mathbf{Z}=\mathbf{0})+\beta_k \mathbf{z}}}, \tag{24}$$

where the new intercept functions $\gamma_k(t|\mathbf{Z} = \mathbf{0})$ again determine the shape of the baseline distribution functions $F_j(t|\mathbf{Z})$. This proportional odds model again requires appropriate constraints on the functions $\gamma_1, \ldots, \gamma_m$ for (3.11) to yield proper survival functions. Although the model (3.11) differs from (3.7) there is no *a priori* reason to prefer one over the

other.

4 Unlinked Regression Models for Current Status Data

In the competing risks and multi-state survival scenarios of Sections 2–3, we avoided the use of simple unlinked regression models for the sub-distribution functions in the former case, and the marginal distribution functions in the latter, since the use of such may not lead to a proper joint distribution function. However, as we have now explored, correspondences between a full data regression model and a multivariate binary regression model for incomplete current status observations are not as straightforward as in the univariate setting, at least when additive effects of the monitoring time and covariates are desired. Further, Jewell, van der Laan and Henneman [13] show that, in the competing risks setting, smooth functionals of the sub-distribution functions can be efficiently estimated—asymptotically—using separate unlinked nonparametric maximum likelihood estimators of the individual sub-distribution functions. The advantage of this approach is that the unlinked estimators are much simpler than the full nonparametric maximum likelihood estimator while they retain consistency. A similar result was established for nonparametric estimators of the marginal distributions for finite multi-state counting processes in van der Laan and Jewell [22]. This suggests that there may be little or no asymptotic precision gained by estimating regression relationships jointly rather than separately, and that the simpler estimators may, in fact, outperform, the more complex simultaneous modeling investigated in Sections 2–3 with small or moderate sample sizes. While this opinion is speculative and remains to be more fully addressed elsewhere, both in theory and simulations, we give a brief outline of this strategy here.

4.1 Competing Risks Models

We continue to use the notation of Section 2. Recall that current status data is represented by (C, Δ), where $\Delta = 0$ if $T \geq C$, $\Delta = j$ if $T < C$ with $J = j$, for $1 \leq j \leq m$. Define the observed binary random variables $\Psi_j = 1$ if $\Delta = j$ and $\Psi_j = 0$ otherwise. Note that

$$E(\Psi_j | C, \mathbf{Z} = \mathbf{z}) = F_j(C; \mathbf{z}), \tag{25}$$

for $1 \leq j \leq m$. Thus, taking each j separately, (4.1) allows construction of a regression model for $F_j(t; \mathbf{z})$ in correspondence with a binary regression model for Ψ_j as for standard univariate current status data. For example, a logistic regression model for Ψ_j with covariates $\mathbf{Z}$ leads to a proportional odds relationship between $\mathbf{Z}$ and $F_j(t; \mathbf{z})$ as in (1.1), the

only difference being that $F_j(\infty; \mathbf{z})$ may be less than 1 so that the corresponding incidence function $\alpha_j(t)$ potentially has a finite limit at ∞.

The advantage to using these separate models is their simplicity, with the consequence that they can be fit using standard software for univariate current status data, leading to semi-parametric estimators $\hat{F}_j(t; \mathbf{z})$ for $j = 1, \ldots, m$ and any t and $\mathbf{z}$. The disadvantage, as previously noted, is that, even though $\hat{F}_j(t; \mathbf{z})$ is non-decreasing in t for any fixed value of $\mathbf{z}$ as desired, $\sum_{j=1}^m \hat{F}_j(t; \mathbf{z})$ may exceed 1 for some values of t and $\mathbf{z}$, violating the requirement that $F(t; \mathbf{z}) = \sum_{j=1}^m F_j(t; \mathbf{z})$ is a distribution function. This, however, may not be a major drawback in large samples as the estimator $\sum_{j=1}^m \hat{F}_j(t; \mathbf{z})$ will consistently estimate the true $F(t; \mathbf{z})$ so long as appropriate semiparametric estimation procedures are used for the separate regression models.

A slight variant on this strategy can be described as follows. First, we use standard univariate current status regression methods to yield an estimator $\hat{F}(t; \mathbf{z})$, based on the observations $(C_i, (\Psi)_i)$ where $\Psi = \sum_{j=1}^m \Psi_j$ indicates only whether the outcome event has occurred by time C without regard to failure type.

Now, for each j, consider the constructed variable $W_j = F(C)\Psi_j$, and note that $E(W_j | C, \Psi = 1) = F_j(C)$. This suggests using current status type regression techniques (that is, isotonic dependence on C and additive linear dependence on $\mathbf{Z}$ with an appropriate link function) for the constructed outcomes $(W_j)_i = \hat{F}(C_i; \mathbf{z}_i)(\Psi_j)_i$ against C_i, using only observations with $(\Psi)_i = 1$, that is, observations where an event of any type has occurred by the monitoring time. This yields estimators $\hat{F}_j(t; \mathbf{z})$ for each j.

While this approach still does not guarantee estimators $\hat{F}_j(t; \mathbf{z})$ which sum to less than 1, this may be somewhat less likely than the first unlinked method since, for each j, the constructed outcomes $(W_j)_i$ are smaller than the respective outcomes $(\Psi_j)_i$ for the previous estimators. In the single sample setting, this approach is related to the full nonparametric maximum likelihood estimator of $F_1, \ldots, F_m$—see Jewell *et al.* [13].

4.2 Multi-State Survival Models

With the notation of Section 3, current status data is given by (C, Φ), where $\Phi = j$ if $T_{j-1} < C \leq T_j$ for $j = 1, \ldots, m+1$, where $T_0 \equiv 0$ and $T_{m+1} \equiv \infty$. In this setting define $\Psi_j = 1$ if $\Phi > j$, and $\Psi_j = 0$ otherwise. Note that

$$E(\Psi_j | C, \mathbf{Z} = \mathbf{z}) = F_j(C|\mathbf{z}), \tag{26}$$

for $1 \leq j \leq m$. Thus, we can separately estimate marginal regression models for $F_j(t|\mathbf{z})$ for each j using univariate current status methods on the data (C, Ψ_j). Again, the advantages of this approach are simplicity, use of standard current status methods only, and direct regression modeling of the marginal distributions, presumably the primary relationships of interest. But once more, although estimates of $F_j(t|\mathbf{z})$ obtained in this way are each distribution functions they are not guaranteed to be stochastically ordered, as required by the structure of the data. Again, this is unlikely to be a serious problem in large samples for similar reasons to those discussed with competing risks data.

Finally, there are variants to this approach similar to that suggested in Section 4.1 for competing risks data. For example, suppose we obtain the estimator $\hat{F}_1(t; \mathbf{z})$ using the data on Ψ_1 as described. Now, consider the constructed variable $W_2 = F_1(C)\Psi_2$, where again it immediately follows that $E(W_2|C, \Psi_1 = 1) = F_2(C)$. As before, this suggests using current status regression techniques for the constructed outcomes $(W_2)_i = \hat{F}_1(C_i|\mathbf{z}_i)(\Psi_2)_i$, against C_i, using only observations with $(\Psi_1)_i = 1$, thereby yielding an estimator $\hat{F}_2(t; \mathbf{z})$. This process then is repeated to yield estimators $\hat{F}_3(t; \mathbf{z}), \hat{F}_4(t; \mathbf{z})$, and so on. Again this approach does not guarantee stochastic ordering of the estimated marginals of F, although it may be more likely since, for each j, the constructed outcomes $(W_j)_i$ are smaller than the respective outcomes $(\Psi_j)_i$ for the previous estimators (and smaller than $\hat{F}_{j-1}(C_i; \mathbf{z}_i)$).

5 Motivating Examples

We briefly describe illustrations of competing risks and multi-state survival data where the need for practical regression models for current status data motivated the development in the earlier sections. In the competing risks case, Krailo and Pike [16] discuss data from the National Center for Health Statistics' Health Examination Survey, originally analysed by McMahon and Worcester [19]. In particular, they focus on the menopausal history of 3,581 female respondents from 1960-1962 who provided cross-sectional information on their age and their menopausal status. For those who had experienced menopause, further retrospective information on the exact age when their periods stopped was deemed unreliable by McMahon and Worcester because of extreme digit preference. Thus, Krailo and Pike [16] concentrated on the simple current status information on menopausal status, in addition to the response on whether menopause had occurred due to an operation or not. Thus natural and operative menopause provide the two causes of 'failure' (here, menopause) in the context of competing risks. Jewell *et al.* [13] analyze this current status data with a nanparametric model. To extend these 'one-sample' models to allow for regression effects requires the kinds of models introduced in Section 2.

This example suggests interesting extensions to simple current status observation of competing risks data. According to MacMahon and Worcester [19], the original data from the Health Examination Survey contained reliable information about the exact age at operative menopause, despite the concerns about information about age at natural menopause. This raises the problem of estimation of regression models for the subdistribution functions F_1 and F_2 in the case where exact times of failures are observed when a failure due to the first risk has occurred before the observation time but where only current status information is available regarding failures due to the second risk. Jewell *et al.* [13] consider this problem in the 'one sample' case.

We now turn briefly to examples of regression based on current status observation of a multi-state survival process, namely the onset and diagnosis of uterine fibroids. The compound 2,3,7,8-tetrachlorodibenzo-*p*-dioxin, commonly known as TCDD or dioxin, is a toxic hydrocarbon and environmental contaminant. It has a half-life of approximately 8 years in humans and, in addition to being a carcinogen, has been shown to disrupt endocrine pathways. On July 10, 1976, an explosion at a chemical plant in Seveso, Italy, exposed local residents to the highest known environmental dioxin levels in a residential area of about 18 km^2 around the plant. A number of health assessments were launched soon after the explosion and many blood samples were collected from residents with sera stored for subsequent analyses. The Seveso Women's Health Study (SWHS) was initiated in 1996, assembling a historical cohort of more than 500 women who were under 40 years of age at the time of the explosion, who were resident in the most heavily exposed areas, and who had sufficient stored sera from the period 1976–1980 available for analysis. Individual level of dioxin exposure was evluated using the stored sera. For a detailed description of the study see Eskenazi *et al.* [9].

Uterine fibroids are noncancerous growths in the uterus, commonly referred to as fibroids. Although uterine fibroids may be present in up to 75% of all women, about a half of these women do not have symptoms. Symptoms, leading to a diagnosis, may develop slowly over a period of several years or rapidly over a period of several months and may include abnormal menstrual bleeding, pelvic pain and pressure and urinary problems. During the period 1996–98 eligible women—still menstruating—in the SWHS were interviewed and received a transvaginal ultrasound, a screening instrument that can detect the presence of fibroids in women without symptoms. Prior diagnosis of fibroids was determined at interview and medical records used to calculate the age at diagnosis. With age as the time scale of interest, all women included in the analysis contributed curent status data on onset of the disease with medical records potentially providing exact ages at diagnosis where this had occurred. If only the prior existence of a diagnosis of fibroids is known, then the data structure corresponds with what is envisioned in Section 3 where the monitoring time corresponds with age at screening. Here, regression effects may focus on dioxin

exposure information although other covaraite effects may also be of sunstantial interest. As in the case of the competing risks example, right-censored information on the age at diagnosis at the time of screening provides an interesting variant to the 'pure' current status form of data structure considered in Section 3. van der Laan *et al.* [23] consider a 'one sample' version of this kind of data structure. Dunson and Baird [7] consider a regression model in this context, with their appropach also applied to the analysis of fibroids data arising from a National Institute of Environmental Health Sciences cross-sectional study of the premenopausal incidence of uterine fibroids. The primary covariate of interest in their regression analysis was race. Young and Jewell [25] compare Dunson and Baird's [7] model to an extension of the approach of van der Laan *et al.* [23] to the regression setting using data examples and simulations.

Multi-state examples occur in quite different contexts than disease progression. For example, in cross-sectional life/sexual history surveys questions are often asked about the number of distinct sexual partners experienced by the respondent by their age at survey. Similarly, employment history questionaires may focus on the number of distint employment (or unemployment) experiences of the respondent. Often, with such data, there may be little or no information on the exact ages where a respondent transitions between 'states' that describe the current cumulative number of partners or experiences. This therefore produces current status data of exactly the sort considered in Section 3. Although such data precludes sudy of association bretween the time spent in various states, there is often still considerable interest in investigating and comparing marginal regression models for times until specified transitions.

6 Discussion

We have considered correspondences between regression models for multinomial outcomes and various multivariate survival models that extend those developed by Doksum and Gasko [6] in a univariate setting. While this suggests some useful regression survival models that can be identified from current status observation, the correspondences are not generally straightforward. This motivates the simpler approach of examining several unlinked univariate regression models as suggested in Section 4. However, there are a wider range of multinomial models that can be considered here so that this should only be considered as a preliminary investigation. Doksum and Gasko [6] also consider correspondences with linear transformation models. It is natural to consider extensions of these ideas to the multivariate setting in which multivariate survival regression models correspond to multivariate binary anlaogues. Space does not permit further discussion of results in this area and details will appear elsewhare. It is important to note that several approaches to multivariate current

status data wityh a common monitoring time have already appeared (Wang and Ding [24], Dunson and Dinse [8], Ding and Wang [5], Jewell, van der Laan and Lei [12]).

References

[1] Agresti, A. (2002). *Categorical Data Analysis*. 2nd ed. Wiley, New York MR1914507

[2] Bennett, S. (1983). Analysis of survival data by the proportional odds model. *Statistics in Medicine* **2**, 273–7.

[3] Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* **34**, 187–220. MR0341758

[4] Crowder, M.J. (2001). *Classical Competing Risks* Chapman & Hall, New York.

[5] Ding, A.A. & Wang, W. (2004). Testing independence for bivariate current status data. *J. Amer. Statist. Assoc.* **99**, 145–55. MR2054294

[6] Doksum, K.A. & Gasko, M. (1990). On a correspondence between models in binary regression analysis and in survival analysis. *International Statistical Review* **58**, 243–52.

[7] Dunson, D.B. & Baird, D.D. (2001). A flexible parametric model for combining current status and age at first diagnosis data. *Biometrics* **57**, 396–403. MR1855672

[8] Dunson, D.B. & Dinse, G.E. (2002). Bayesian models for multivariate current status data with informative censoring. *Biometrics* **58**, 79–88. MR1891046

[9] Eskenazi, B., Mocarelli, P., Warner, M., Samuels, S. Vercellini, P., Olive, D., Needham, L., Patterson, D. & 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.

[10] Jewell, N.P. & van der Laan, M. (1995). Generalizations of current status data with applications. *Lifetime Data Analysis* **1**, 101–9. MR1425898

[11] Jewell, N.P. & van der Laan, M. (1997). Singly and doubly censored current status data with extensions to multi-state counting processes. In *Proceedings of First Seattle Conference in Biostatistics* Lin, D-Y. ed., Springer Verlag, 171–84.

[12] Jewell, N.P., van der Laan, M. & X. Lei (2005). Bivariate current status data with univariate monitoring times. *Biometrika* **92**.

[13] Jewell, N.P., van der Laan, M. & Henneman, T. (2003). Nonparametric estimation from current status data with competing risks. *Biometrika* **90**, 183–97. MR1966559

[14] Jewell, N.P. & 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, Elsevier, Amsterdam. MR2065792

[15] Kalbfleisch, J.D. & Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. Wiley, New York. MR1924807

[16] Krailo, M.D. & Pike, M.C. (1983). Estimation of the distribution of age at natural menopause from prevalence data. *Am. J. Epidemiol.* **117**, 356–61.

[17] Larson, M.G. & Dinse, G.E. (1985). A mixture model for the regression analysis of competing risks data. *Applied Statistics* **34**, 201–11. MR0827668

[18] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*. 2nd ed. Chapman & Hall, New York.

[19] MacMahon, B. & Worcester, J. (1966). Age at menopause, United States 1960–1962. *National Center for Health Statistics; Vital and Health Statistics, Series 11: Data from the National Health Survey, no. 19* Washington, DC: DHEW Publication no. (HSM) 66–1000.

[20] Rossini, A. & Tsiatis, A.A. (1996). A semiparametric proportional odds regression model for the analysis of current status data. *J. Amer. Statist. Assoc.* **91**, 713–21. MR1395738

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

[22] van der Laan, M. & 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. MR1983540

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

[24] Wang, W. & Ding, A.A.(2000). On assessing the association for bivariate current status data. *Biometrika* **87**, 879–93. MR1813981

[25] Young, J.G. & Jewell, N.P. (2006). Regression analysis of onset and diagnosis distributions. Preprint.