

Semiparametric Regression Models for Repeated Events with Random Effects and Measurement Error

Wenxin JIANG, Bruce W. TURNBULL and Larry C. CLARK ¹

Abstract

Statistical methodology is presented for the regression analysis of multiple events in the presence of random effects and measurement error. Omitted covariates are modeled as random effects. Our approach to parameter estimation and significance testing is to start with a naive model of semi-parametric Poisson process regression, and then to adjust for random effects and any possible covariate measurement error. We illustrate the techniques with data from a randomized clinical trial for the prevention of recurrent skin tumors.

KEY WORDS: Consistency; Cox model; Estimating equations; Frailty; Measurement error; Omitted covariates; Point process; Poisson regression; Proportional intensities; Robust estimator; Selenium; Skin cancer; Specification analysis; Unobserved heterogeneity; Validation data.

1. INTRODUCTION

The research in this paper was motivated by some of the statistical discussions leading up to the publication of the somewhat controversial findings of the “Nutritional Prevention of Cancer” (NPC) trial — Clark et al. (1996). This trial, begun in 1983, studied the long-term safety and efficacy of a daily $200\mu\text{g}$ nutritional supplement of selenium (Se) for the prevention of cancer. This was a double-blind, placebo-controlled randomized clinical trial with 1312 patients accrued and followed for up to about ten years. A number of endpoints were considered, but here we shall concentrate on one of the two primary endpoints — namely squamous cell carcinoma (SCC) of the skin. The results for this endpoint are of particular interest because Clark et al. (1996) found a negative (but not statistically significant, $P = 0.15$) effect of selenium (Se) supplementation. This was opposite to previous

¹Wenxin Jiang is Assistant Professor, Department of Statistics, Northwestern University, Evanston, IL 60208. Bruce W. Turnbull is Professor, School of Operations Research and Department of Statistical Science, 227 Rhodes Hall, Cornell University, Ithaca, NY 14853. Larry C. Clark is Associate Professor, Arizona Cancer Center, University of Arizona, Tucson AZ 85724. The first author was supported in part by NSF grant DMS-9505799, the second author by NIH grant R01 CA 66218, and the third author by NIH grant R01 CA 49764. The authors are grateful to a referee, who made several very helpful suggestions.

expectations, and contrasted sharply with findings of highly significant positive benefits of the selenium supplementation in preventing a number of other types of cancers. However, for the SCC endpoint, the original analysis presented in Clark et al. (1996) considered only the time to *first* occurrence of an SCC in each subject. In fact, subjects could experience multiple recurrences of SCCs over time. One question posed was how to incorporate this information which would presumably lead to a more sensitive inference concerning the effect of Se supplementation. Complicating the situation was the fact that the subjects were highly heterogeneous. Various demographic, behavioral and medical “baseline” variables had been measured upon entry to the study for each patient, but there was concern that not all risk factors had been identified. It was also recognized that some of these baseline variables were subject to biological variability and measurement error, especially blood biochemical levels, such as plasma Se status, which, according to some earlier epidemiologic evidence, is an important prognostic risk factor for SCC.

Situations where individual subjects (or units) may experience repeated events over time occur not only in medical trials, — studies of epileptic seizures, asthmatic attacks, infections, for example, — but also in engineering studies of reliability of repairable systems (e.g., Ascher and Feingold, 1984), and in “duration analysis” studies in economics and sociology (e.g., Allison, 1984). We will suppose that subjects are heterogeneous and some covariates are available from each subject. The event times for each subject can be viewed as a realization of a stochastic point process. To analyze such data, a variety of parametric point process regression models have been proposed — a good discussion appears in Lawless (1987). A more flexible approach is to use a semiparametric model, in which the functional form of the baseline intensity function is unspecified. There are four such approaches which may all be considered generalizations of the Cox (1972) model for survival data. These are: the counting process formulation of Andersen and Gill (1982); the marginal approach of Wei, Lin and Weissfeld (1989); the conditional approach of Prentice, Williams and Peterson (1981); and the cumulative mean function approach of Lawless and Nadeau (1995). Statistical software to implement parametric and semi-parametric methods are available in standard computer packages, e.g., STATA 5.0 (StataCorp 1997) and Splus 3.3 (Statistical Science Inc. 1995).

In the next two sections we describe our approach which is to use, as a basis for inference, estimating equations derived from a naive semiparametric Poisson process regression model in which the presence of random effects, omitted covariates, and measurement error are ignored. These “naive” estimates of the regression parameters and of the baseline intensity function, along with their estimated variances (which may be computed by the standard software mentioned above), must then be adjusted to account for the misspecification of the

model. In Section 4, we consider the special case when there is no measurement error; the point estimates are unchanged, but their variances must be estimated using the so-called “sandwich” formula. This case was treated by Lawless and Nadeau (1995). We provide asymptotic expressions of these variances and show that they are determined by the first two moments of the point process. In Section 5, we consider the special case of overdispersion in a Poisson process regression model and use it to examine the inflation in the variance that can be caused by the presence of random effects and omitted covariates, and to illustrate the evaluation of the relative efficiencies of our estimators. When measurement error is also present, the naive point estimators are no longer consistent but, with a model for the measurement error process, they can be adjusted to obtain consistent estimators. Also, the estimated variances of these adjusted estimators are then given by a “double sandwich” formula. This is all described in Section 6. The methods are applied to the NPC data in Section 7. Finally we make some concluding remarks concerning assumptions.

For the special case of survival analysis, where subjects can experience at most a single event during their followup time, a number of researchers have examined the problem of covariate measurement error in a partial likelihood analysis using Cox’s (1972) model. Prentice (1982) and Pepe et al. (1989) considered the induced hazard function conditional on the observed covariate instead of its true value. Nakamura (1992) and Buzas (1998) used procedures based on constructing unbiased or approximately unbiased estimating equations from the partial likelihood score equations. The methods of Raboud (1991) and Raboud et al. (1993) were based on examining the root of the asymptotic score equation of the naive partial likelihood. There is also an extensive literature on misspecification of the Cox model due to the presence of frailties, omitted or mismodeled covariates, see Lin and Wei (1989), Gail, Wieand and Piantadosi (1984), Lagakos and Shoenfeld (1984), and a recent review by Keiding, Andersen and Klein (1997). The effects of random subject heterogeneity and covariate measurement error on fully parametric analyses of recurrent events, based on a time-homogeneous Poisson process with gamma mixing, have been discussed by Turnbull, Jiang and Clark (1997), of which the present paper may be regarded as a semi-parametric analog.

2. A GENERAL REGRESSION MODEL FOR RECURRENT EVENTS

We consider a discrete time scale $\mathcal{K} = \{1, \dots, K\}$, measured in “days” say, and define response Y_{ik} to be the number of events for subject i ($1 \leq i \leq n$) observed on day k , ($k \in \mathcal{K}$). We denote by Z_{ik} the observed value of the covariate (a p -vector in general) for patient i on day k , and H_{ik} is a 0-1 variable indicating if subject i is “at risk”, i.e., uncensored on

day k . Note that $Y_{ik} \geq 1$ only when $H_{ik} = 1$. In general, each observed covariate Z_{ik} may be measured with error, or is a surrogate of some “true” covariate X_{ik} , usually unobserved. In the NPC trial example of the preceding section, the time scale is the elapsed number of days from date of entry into the study, and the response Y_{ik} is the number of SCCs (usually 0 or 1) recorded on day k for subject i . Many covariates could be considered but suppose we concentrate here on $p = 2$ of the most important, namely treatment assignment (Se or placebo) and baseline plasma Se status. The latter is subject to measurement error, but the former is presumably accurate. In this case, the covariates are not time varying and so $\{Z_{ik}\}$ (and $\{X_{ik}\}$) do not depend on k . There were $n = 1312$ subjects and the maximum followup time was $K = 4618$ days. For example, one patient had four recurrent SCCs recorded on respective days 178, 286, 549, and 1018, and was followed for a total of 2268 days, when the study ended and he was censored. Thus, for him, all $Y_{ik} = 0$ except $Y_{ik} = 1$ for $k = 178, 286, 549, 1018$; also $H_{ik} = 1$ for $1 \leq k \leq 2268$ and $H_{ik} = 0$ for $2269 \leq k \leq 4618$.

We also postulate an unobserved positive real-valued variable ψ_{ik} which represents a random effect which modifies response for patient i on day k . This random effect or “frailty” factor is introduced to model the effect of “unexplained heterogeneity” of patients perhaps induced by omitting covariates, unused, unmeasured or undreamed of. We introduce the notation Y_i^* to represent the complete response history $\{Y_{ik}; 1 \leq k \leq K\}$ of subject i , ($1 \leq i \leq n$), and similarly $H_i^*, X_i^*, Z_i^*, \psi_i^*$. Finally we assume a setup where the concatenated vectors $\{Y_i^*, H_i^*, X_i^*, Z_i^*, \psi_i^*\}, i = 1, 2, \dots, n$ may be considered independent and identically distributed (i.i.d.). A general regression model for recurrent events is the following:

Multiplicative Mean (MM) Model:

Conditional on $\{H_i^*, X_i^*, Z_i^*, \psi_i^*\}$, the expectations of responses Y_{ik} are given by $H_{ik}\psi_{ik}\Lambda_k \exp(X_{ik}'\beta)$, where we assume $E\{\psi_{ik}|X_i^*, Z_i^*, H_i^*\} = 1$, for all $i = 1, 2, \dots, n, k \in \mathcal{K}$.

Here β is a p -vector of regression coefficients, Λ_k is the discrete baseline intensity on day k . The baseline cumulative intensity function is $\Lambda_0(t) = \sum_{k \leq t} \Lambda_k, t \in \mathcal{K}$.

The MM model is a quite general one with essentially two modeling assumptions — one involves the structure of the frailties, the other concerns how the covariates affect the mean response rate. The first states that the conditional means of the frailties $E\{\psi_{ik}|X_i^*, Z_i^*, H_i^*\}$ are constant for all $i = 1, 2, \dots, n, k \in \mathcal{K}$. Without loss of generality, this constant is taken as unity. Without at least such an assumption, the frailties would be arbitrary and there would be no restriction on the conditional mean responses. The assumption implies the loglinear model $E\{Y_{ik}|X_i^*, Z_i^*, H_i^*\} = H_{ik}\Lambda_k \exp(X_{ik}'\beta)$. However, it is advantageous to state the MM model at the level of conditioning on the ψ_{ik} 's, since a simpler probability model can

often be expressed at this level of full conditioning. An example of this in continuous time is given in Lawless (1987), where the responses follow a Poisson process with Weibull intensity upon conditioning on a time-invariant random effect ψ_i , which follows a gamma distribution with unit mean. The second modeling assumption is that a loglinear link function is used to relate the effect of covariates to the event rates. Other particular link functions could be specified instead (see Lawless and Nadeau 1995); however, the loglinear one is often a natural and convenient choice for count data (McCullagh and Nelder, 1989, Sec.2.2), and commonly used in the survival analysis or repeated events literature (e.g., Lawless 1987, and Wei, Lin and Weissfeld 1989). Sometimes log-linearity will be reasonable after a suitable transformation of the covariates (see, e.g., Andersen et al. 1993, Sec. VII.3.2) even if not so originally. However, many of our results apply to general link functions — we will comment on this further in Section 8.

As a basis for inference in the MM model, we propose to use a “naive” or “working” likelihood function obtained by additionally taking the Y_{ik} to be (conditionally) independent and Poisson distributed and by neglecting the presence of random effects and measurement errors. Thus all the ψ_{ik} ’s are taken as 1, and X_{ik} ’s are replaced with Z_{ik} ’s.

The resulting misspecified or “naive” log likelihood function $R = R(s)$ is, up to a constant in arguments $s = (m', b)'$ where $m = (m_1, \dots, m_K)'$, given by

$$R = \sum_{i=1}^n \log \prod_{k=1}^K \{(m_k e^{Z'_{ik} b})^{H_{ik} Y_{ik}} \exp(-H_{ik} m_k e^{Z'_{ik} b})\}.$$

Here, as the argument of R , we have replaced the true parameter $\theta = (\Lambda', \beta)'$ where $\Lambda = (\Lambda_1, \dots, \Lambda_K)'$, with $s = (m', b)'$, to emphasize the fact that we are using a misspecified model in which the parameters may not have the same interpretation. We note that R is a sum of n i.i.d. copies of the function $\rho = \rho(s)$, given by

$$\rho = \sum_k H_k \{Y_k Z'_k b + Y_k \log m_k - m_k e^{Z'_k b}\}$$

where we have suppressed the index i for the i.i.d. subjects.

For fixed b , R is maximized at $m_k = (\sum_i Y_{ik}) (\sum_i H_{ik} e^{Z'_{ik} b})^{-1}$. Substituting this value of m_k into R , we find that the function R on the “ridge” is just Cox’s log partial likelihood function \mathcal{L} of the argument b , up to some constant, where $\mathcal{L}(b) = \log \prod_i \prod_k (e^{Z'_{ik} b} / \sum_j H_{jk} e^{Z'_{jk} b})^{Y_{ik}}$. Similar arguments in counting process theory can be seen in Andersen et al. (1993, Sec.VII.2.1, page 482). Hence R is maximized by the “naive” maximum likelihood estimates (MLEs) $b = \hat{b}$ and $m = \hat{m}$, where

$$\hat{b} = \arg \max \mathcal{L}(b) \quad \text{and} \quad \hat{m}_k = \left(\sum_i Y_{ik} \right) \left(\sum_i H_{ik} e^{Z'_{ik} \hat{b}} \right)^{-1}. \quad (1)$$

Here the $\hat{m}_k = (\hat{m})_k$, $k = 1, \dots, K$ form a discrete version of the Nelson-Aalen estimates (see Andersen et al. 1993, Sec.VII.2), which lead to a naive estimator of the baseline cumulative intensity function $\hat{m}_0(t) = \sum_{k \leq t} \hat{m}_k$, $t \in \mathcal{K}$.

We now examine the consistency and asymptotic variance of these naive estimators under the MM model.

3. ASYMPTOTIC PROPERTIES OF THE NAIVE ESTIMATORS

A natural question is to ask how appropriate are the naive estimators and how they and their variances should be adjusted to account for the misspecified features. We base our answers on the asymptotic properties of such naive likelihood estimates, — see Huber (1967), White (1994), Jiang (1996), Turnbull et al. (1997). The principal results can be summarized as follows:

Proposition 1. Suppose W_i $i = 1, \dots, n$ are i.i.d. copies of W , where W has a probability distribution $P_W^{(\theta)}$ with parameter θ lying in some space Θ . Let $R_n(s) = \sum_{i=1}^n \rho(W_i; s)$, which could be a naive log-likelihood function with argument s . Under some regularity conditions, we have:

- A. $\hat{s}_n \equiv \arg \max_s R_n(s)$ is strongly consistent to $s^0(\theta) = \arg \max_s E_\theta \rho(W; s)$, and $\sqrt{n}(\hat{s}_n - s^0(\theta)) \xrightarrow{\mathcal{D}} N(0, I^{-1}VI^{-1})$ where $I = -E\{\nabla^{\otimes 2} \rho(W; s)\} \Big|_{s^0(\theta)}$ and $V = E\{\nabla \rho(W; s)\}^{\otimes 2} \Big|_{s^0(\theta)}$;
- B. An inverse function $(s^0)^{-1}$ exists, $\hat{\theta}_n = (s^0)^{-1}(\hat{s}_n)$ is strongly consistent to the original parameter θ , and $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} N(0, D'I^{-1}VI^{-1}D)$ with $D = \nabla(s^0)^{-1} \Big|_{s^0(\theta)}$;
- C. $s = s^0(\theta)$ satisfies the estimating equation

$$E_\theta \nabla \rho(W; s) = 0. \tag{2}$$

In the notation above, E or E_θ represents taking expectation over W , based on the probability distribution $P_W^{(\theta)}$ associated with the true parameter θ . The symbol ∇ represents the column vector of partial derivative operators (the gradient) with respect to the argument s . The operation $\otimes 2$ of a column vector v denotes $v^{\otimes 2} = vv'$. Hence $\nabla^{\otimes 2}$ denotes the matrix of second derivatives (Hessian). Essentially the regularity conditions that are needed in Proposition 1 can be listed as follows:

1. As $n \rightarrow \infty$, the average naive likelihood $n^{-1}R_n(s)$ is uniformly convergent to its expectation on a compact set Ξ with probability one; and the limit function $(E\{n^{-1}R_n(s)\})$

or $E\rho(W; s)$) is sufficiently regular, in the sense that it is finite and second order continuously differentiable, with the derivatives commuting with the expectation;

2. The limit function $E\{n^{-1}R_n(s)\}$ has a unique global maximizer $s^0(\theta)$ in the interior of Ξ ;
3. The matrices I and V are finite and non-singular, so that the asymptotic variance of $\sqrt{n}\hat{s}_n$ is well-defined and non-singular;
4. $s^0(\cdot) : \Theta \mapsto s_0(\Theta)$ is a \mathcal{C}^1 -diffeomorphism.

A variety of more primitive conditions to ensure the uniform convergence with a regular limit (Condition 1) can be found in White (1994, Appendix 2), Jiang (1996), and Turnbull et al. (1997, Appendix).

In the following discussion, when convenient, we may omit the subscript n for \hat{s}_n , $\hat{\theta}_n$ and R_n , and denote the limit of the naive MLE $s^0(\theta)$ as $s(\theta)$ or simply s , which we sometimes refer to as the “naive” parameter. The relation $s^0(\theta)$ expresses the naive parameter in terms of the original parameter, which we term as the “bridge relation”. There are situations when the expectation in equation (2) depends on parameters in addition to θ which are not estimable from the data W_i , $i = 1, \dots, n$. Such a situation can occur, for example, if there is covariate measurement error involved. If these extra parameters cannot be treated as known, then they will need to be estimated from an auxiliary or “validation” data set, see e.g., Carroll et al. (1995, Sec. 1.4).

Note that the naive MLE \hat{s} is in general **inconsistent** for the original parameter θ , but instead consistently estimates the naive parameter $s(\theta)$. When $s(\theta)$ is invertible, we can form a consistent estimator $\hat{\theta}$ for θ by inverting the relation $s(\theta)$, as suggested by result (B) of the proposition. The general strategy therefore is to attempt to find the bridge relation $s(\theta)$ using result (A) or (C) of the proposition and then, when the relation is invertible, to use result (B) to obtain an “adjusted” estimator of θ .

The asymptotic normality of \hat{s} and $\hat{\theta}$ from (A) and (B) of Proposition 1 shows how standard errors and test statistics can be constructed. The “naive” asymptotic variance (matrix) of \hat{s} that ignores the model misspecification is $(nI)^{-1}$, the inverted naive information matrix. The correct asymptotic variance of \hat{s} is $n^{-1}I^{-1}VI^{-1}$, the so-called “sandwich formula” — Huber (1967), Carroll et al. (1995, page 263). We will later refer to this as the “robust variance” since it is valid, in the sense of yielding consistent variance estimators, without full specification of a probability model for the data. The asymptotic variance of the adjusted estimate $\hat{\theta}$ is thus given by $n^{-1}D'I^{-1}VI^{-1}D$, which we term as the “double-sandwich formula”. If the expectations in I and V are not available then quantities based on sample

averages can be used in the usual way, i.e., for I use $\hat{I} = -n^{-1} \sum_i \{\nabla^{\otimes 2} \rho(W_i; s)\} |_{\hat{s}}$, for V use $\hat{V} = n^{-1} \sum_i \{\nabla \rho(W_i; s)\}^{\otimes 2} |_{\hat{s}}$. To test the significance of a particular component of the parameter vector, the j th say, using the Wald method, the naive test statistic that ignores model misspecification would be the \mathcal{Z} -value: $\mathcal{Z}_N = \hat{s}_j / \sqrt{Avar_N(\hat{s}_j)}$ where $Avar_N(\hat{s}_j)$ is the j th diagonal element of $(nI)^{-1}$. The correct \mathcal{Z} -statistic is: $\mathcal{Z}_{Adj.} = \hat{\theta}_j / \sqrt{Avar(\hat{\theta}_j)}$ where $Avar(\hat{\theta}_j)$ is the j th diagonal element of $n^{-1} D' I^{-1} V I^{-1} D$.

In the context of the recurrent event regression models described in Section 2, W is the collection of random variables $\{H_k, Y_k, Z_k; k \in \mathcal{K}\}$, $\theta = (\Lambda', \beta)'$, and $s = (m', b)'$. (Here and later, for notational convenience, we may suppress the subject-specific index i due to the i.i.d. property.) It is the goal of this paper to investigate the relationship between the quantities obtained from the naive and adjusted analysis, including the parameter estimates, asymptotic variances and \mathcal{Z} -values. The first task is to find the limit $s^0(\theta)$ of the naive MLE.

By using (2) and first conditioning on H^* , X^* and Z^* , from the MM model we directly obtain the following “bridge” equations:

$$m_k E[H_k e^{Z_k' b}] = \Lambda_k E[H_k e^{X_k' \beta}], \quad k = 1, \dots, K \quad (3)$$

$$\sum_{k=1}^K m_k E[H_k Z_k e^{Z_k' b}] = \sum_{k=1}^K \Lambda_k E[H_k Z_k e^{X_k' \beta}]. \quad (4)$$

Here, to ensure the existence of the expectations, we assume the existence of the moment generating functions of the X_k 's and Z_k 's on \mathfrak{R}^p and assume that the first and second order derivatives commute with the expectations. These hold if the probability distributions of the X_k 's and Z_k 's have sufficiently thin tails (e.g., bounded or normal). Note that here in (3,4) and later we are using m_k and b to denote the large-sample limits of the naive estimators \hat{m}_k and \hat{b} , respectively. Similarly we use $m_0(t) = \sum_{k \leq t} m_k$ to denote the large-sample limit of the naive estimator $\hat{m}_0(t) = \sum_{k \leq t} \hat{m}_k$ of the baseline cumulative intensity function. We cannot solve (3,4) explicitly without making further modeling assumptions concerning the measurement error structure of (X_k, Z_k) . The simplest situation is when there is no measurement error, and the only complication is the existence of random effects/omitted covariates. It is this situation that we examine first.

4. RANDOM EFFECTS

When there is no covariate measurement error, X_k equals Z_k for all $k = 1, \dots, K$, and it is clear that there is a trivial solution to (3,4), namely

$$b = \beta, \text{ and } m = \Lambda.$$

The solution is actually unique under the non-degeneracy conditions for the random variables H_k and Z_k , $k \in \mathcal{K}$, namely: (i) For all $k = 1, \dots, K$, $P(H_k = 1) > 0$; (ii) For all k , $\Lambda_k > 0$; and (iii) There do not exist constant p -vector a and scalar c such that $Z_k' a + c = 0$ for all k with probability one. The uniqueness can then be easily shown by first solving (3) for m_k as a function of b and substituting into (4), so that (4) now only has b as an argument. The difference of the two sides of (4) can then be recognized as the gradient of a concave scalar function of b , provided that (i)-(iii) hold.

In the notation of the previous section, the bridge relation is the trivial one $s^0(\theta) = \theta$. This agrees with the findings of Lawless and Nadeau (1995) who showed that estimates derived under a Poisson process regression model were consistent as long as the mean event rates were correctly specified (and the observation process is independent of the event process).

We now turn to the problem of estimating the asymptotic variances of the consistent estimates \hat{b} and \hat{m} . Since the naive MLEs are consistent for the original parameters, D is the identity matrix, and in part (B) of Proposition 1, the double sandwich formula for their asymptotic variance reduces to the usual sandwich formula: $Avar(\hat{\theta}) = n^{-1}I^{-1}VI^{-1}$ for $\hat{\theta} = (\hat{m}', \hat{b}')$, where $I = -E(\nabla^{\otimes 2}\rho)$ and $V = E(\nabla\rho)^{\otimes 2}$. To simplify the notation, we define scalars $\lambda_k = m_k e^{Z_k' b}$ and $\mu_k = H_k \lambda_k$ and note that $\mu_k = E\{Y_k | Z^*, H^*\}$. We also define $\Gamma_{jk} = Cov\{Y_j, Y_k | Z^*, H^*\}$, for each $j, k = 1, \dots, K$. Denote $\mathcal{D}_k = \nabla \log \lambda_k$ as a $(K + p)$ -vector, and $\zeta_k = Z_k - E(Z_k \mu_k) / E\mu_k$ as a p -vector with components $(\zeta_{k1}, \dots, \zeta_{kp})'$, for each $k = 1, \dots, K$. Then $\rho = \sum_{k=1}^K H_k (Y_k \log \lambda_k - \lambda_k)$.

Proposition 2.

(i) The naive variance $Avar_N(\hat{\theta}) = n^{-1}I^{-1}$ and the robust variance $Avar(\hat{\theta}) = n^{-1}I^{-1}VI^{-1}$ can be calculated from the mean function $\{\mu_k\}$ and the covariance function $\{\Gamma_{jk}\}$ using

$$I = \sum_{k=1}^K E(\mu_k \mathcal{D}_k \mathcal{D}_k') \text{ and } V = \sum_{j=1}^K \sum_{k=1}^K E(\Gamma_{jk} \mathcal{D}_j \mathcal{D}_k'); \quad (5)$$

(ii) The naive variance $Avar_N(\hat{b}) = n^{-1}\mathcal{I}^{-1}$ and the robust variance $Avar(\hat{b}) = n^{-1}\mathcal{I}^{-1}\mathcal{V}\mathcal{I}^{-1}$ can be calculated from

$$\mathcal{I} = \sum_{k=1}^K E(\mu_k \zeta_k \zeta_k') \text{ and } \mathcal{V} = \sum_{j=1}^K \sum_{k=1}^K E(\Gamma_{jk} \zeta_j \zeta_k'). \quad (6)$$

Proof. The result (i) follows directly from the definitions of I and V . Result (ii) involves matrix manipulations which we outline as follows: The matrix I is a $(K + p) \times (K + p)$ matrix such that the matrix $nAvar_N(\hat{b})$ is given by the lower-right $p \times p$ sub-matrix of I^{-1} .

When computed, this comes out to be \mathcal{I}^{-1} as defined in (6). Similarly the matrix $nAvar_N(\hat{b})$ is given by the lower-right $p \times p$ sub-matrix of $I^{-1}VI^{-1}$, which comes out as $\mathcal{I}^{-1}\mathcal{V}\mathcal{I}^{-1}$. This proves (ii).

Expressions for estimators of the asymptotic variances and covariances of the estimates of the regression coefficients and the baseline cumulative intensities are given in Appendix A. They are equivalent to expressions given in Appendix A.2 of Lawless and Nadeau (1995), but in a form directly computable from the data and the naive MLEs.

Proposition 2 enables us to perform a number of important tasks, provided we model, in addition to the first moment from the MM assumption, the second moments $\{\Gamma_{jk}\}$. These include: (i) comparison of the naive asymptotic variance and the robust asymptotic variance; (ii) derivation of formulae for sample size planning in randomized clinical trials; (iii) evaluation of the efficiency of the estimator \hat{b} from the current procedure, when some particular interesting departure from the naive Poisson process regression model can be assumed. In the following section, we will illustrate these under a special case when the true model is assumed to be a Poisson process with overdispersion. In principle, Proposition 2 can be used to carry out analogous calculations for other models, such as ones with AR(1) or exchangeable correlations, although analytical details are more difficult.

5. POISSON PROCESS WITH OVERDISPERSION

In this section we suppose that, conditional on (Z^*, H^*, ψ^*) , the responses Y_k ($k = 1, \dots, K$) are independent and Poisson distributed with mean $\psi\mu_k$. The frailty factor $\psi_k = \psi$ is a random variable, constant over time in each individual, with mean one and constant variance κ . As in the previous section, there is no measurement error, X_k equals Z_k for all $k = 1, \dots, K$, and $b = \beta$, $m = \Lambda$. For this model, by first conditioning on ψ^* , Z^* and H^* , we obtain the covariance function $\Gamma_{jk} = Cov(Y_j, Y_k | Z^*, H^*) = \delta_{jk}\mu_k + \kappa\mu_j\mu_k$, for $j, k = 1, \dots, K$, where δ_{jk} denotes the Kronecker delta. The overdispersed Poisson model is relatively simple, and has been quite commonly used in repeated events literature — e.g., Lawless (1987), Abu-Libdeh et al. (1990), Cook (1995), Turnbull et al. (1997).

First we investigate the relation between the naive variance $Avar_N(\hat{b})$ and the robust variance $Avar(\hat{b})$. In this situation, we claim that $Avar(\hat{b}_j) \geq Avar_N(\hat{b}_j)$, for all $j = 1, \dots, p$, and the difference is proportional to the overdispersion variance κ , indicating that the robust variance automatically takes into account the extra variability due to neglected random effects. To show this, we substitute $\Gamma_{jk} = \delta_{jk}\mu_k + \kappa\mu_j\mu_k$ in (6) to obtain $(\mathcal{V} - \mathcal{I})_{rl} = \kappa E(\sum_j \mu_j \zeta_{jr})(\sum_k \mu_k \zeta_{kl})$. Hence the matrix $(\mathcal{V} - \mathcal{I})$ is positive semi-definite and proportional to κ . It follows that the matrix $n^{-1}\mathcal{I}^{-1}(\mathcal{V} - \mathcal{I})\mathcal{I}^{-1} = Avar(\hat{b}) - Avar_N(\hat{b})$, is also positive semi-definite and proportional to κ , showing the result. A similar result can be shown in a similar

fashion for the asymptotic variances of the baseline cumulative intensity estimators, that is $Avar\{\hat{m}_0(t)\} \geq Avar_N\{\hat{m}_0(t)\}$ for all $t = 1, \dots, K$, and the difference is proportional to κ .

Now we illustrate the computation of the robust variance $Avar(\hat{b})$ for the special case of a single binary covariate, constant over time: $Z_k \equiv Z$ and $Z = 0$ or 1 with equal probability 0.5 . This situation occurs, for example, when comparing two treatments in a randomized clinical trial where Z is the treatment assignment indicator. In this case, an expression of the robust variance is useful in sample size planning. In general, the robust variance is dependent on the following four quantities: the treatment effect parameter $\beta (= b)$; the baseline discrete intensity $\Lambda_k (= m_k)$; the covariance function $\Gamma_{jk} = Cov(Y_j, Y_k | H^*, Z)$; and the probability of being followed $P_{Zk} = P(H_k = 1 | Z)$; for $j, k \in \mathcal{K}$ and $Z = 0$ or 1 . Consider the situation when we follow the subjects continuously over time and use the time of randomization (treatment assignment) as the origin of the time axis. Then, for a given subject, the information contained in H^* is the same as in the follow-up time T , say, which is the largest time k in \mathcal{K} at which a subject is followed. We treat T as a random variable. In this notation, we have simply $P_{Zk} = P(T \geq k | Z)$. A general formula of $Avar(\hat{b})$ can be written down directly by using the expression of \mathcal{I} and \mathcal{V} in (6) in terms of b , m_k , $\Gamma_{jk}^Z = E(\Gamma_{jk} | Z)$, say, and P_{Zk} , for $j, k \in \mathcal{K}$ and $Z = 0$ or 1 . The result is summarized below:

Proposition 3. Suppose $Z_k \equiv Z$, and $Z=0, 1$ with equal probability 0.5 . We use the above notation and assume $P_{0k} > 0$ and $P_{1k} > 0$ for all $k = 1, \dots, K$. Then, we can express $Avar(\hat{b}) = n^{-1}\mathcal{I}^{-1}\mathcal{V}\mathcal{I}^{-1}$, where

$$\begin{aligned} \mathcal{I} &= \frac{1}{2} \sum_{k=1}^K m_k \left(\frac{P_{0k}P_{1k}e^b}{P_{0k} + P_{1k}e^b} \right) \quad \text{and} \\ \mathcal{V} &= \frac{1}{2} \sum_{j=1}^K \sum_{k=1}^K \left\{ \frac{\Gamma_{jk}^0(P_{1j}e^b)(P_{1k}e^b) + \Gamma_{jk}^1 P_{0j}P_{0k}}{(P_{0j} + P_{1j}e^b)(P_{0k} + P_{1k}e^b)} \right\}. \end{aligned}$$

In the following, we consider the special case when the follow-up process H^* , or equivalently the follow-up time T , are independent of the randomized treatment indicator Z , and responses follow the overdispersed Poisson process model where $\Gamma_{jk} = \delta_{jk}\mu_k + \kappa\mu_j\mu_k$ with $\mu_k = H_k m_k e^{Zb}$. Then $P_{Zk} = EH_k$, $\Gamma_{jk}^Z = E(\Gamma_{jk} | Z) = \delta_{jk}(EH_k)m_k e^{Zb} + \kappa(EH_j H_k)m_j m_k e^{2Zb}$. Define $\sum_{k=1}^K m_k H_k = \mathcal{T}$, the total cumulative intensity for a subject with $Z = 0$ and with follow-up history H^* . Note that for continuous follow-up from time 1 to T , $\mathcal{T} = \sum_{k=1}^T m_k = m_0(T)$, which is the baseline cumulative intensity (i.e., for a subject with $Z = 0$) at end of the follow-up period. Using Proposition 3 we obtain:

$$\mathcal{I} = \frac{E\mathcal{T}}{2} \frac{e^b}{1 + e^b} \quad \text{and} \quad \mathcal{V} = \mathcal{I} + \kappa E\mathcal{T}^2 \left(\frac{e^b}{1 + e^b} \right)^2.$$

From these we immediately obtain:

$$Avar_N(\hat{\beta}) = (n\mathcal{I})^{-1} = \frac{2}{n} \frac{1}{ET} + \frac{2}{n} \frac{1}{ETe^\beta}$$

(Note that $b = \beta$ and $\hat{\beta} = \hat{b}$ at present). The robust asymptotic variance, on the other hand, is obtained from $n^{-1}\mathcal{I}^{-1}\mathcal{V}\mathcal{I}^{-1}$ as

$$Avar(\hat{\beta}) = \frac{2}{n} \frac{1}{ET} + \frac{2}{n} \frac{1}{ETe^\beta} + \frac{4\kappa}{n} \frac{ET^2}{(ET)^2}. \quad (7)$$

The formula (7) can be used to plan sample sizes for a test of treatment equality with given size and power. Of course, the expression (7) depends on unknown quantities which can only be estimated once the data become available. However preliminary rough estimates may be available from pilot studies. Naturally, the main implication is that the inflation of $Avar(\hat{\beta})$ due to the extra variance component κ , leads to an increase in the required sample size, when compared to that based on the usual Poisson process regression model. This is consistent with previous findings by Cook (1995) for time-homogeneous Poisson processes with overdispersion.

Efficiency properties. Adjusted estimates based on misspecified models and derived using the methods of Section 3, although \sqrt{n} -consistent and asymptotically normal, are generally inefficient. If we are willing to specify fully a particular probability model as the “correct” one, more efficient estimators can be constructed. For example, starting with our \sqrt{n} -consistent adjusted estimator, we can employ a one-step efficientization method (see, e.g., Le Cam 1956, or White 1994, page 137). Alternatively, one can use maximum likelihood. There is the usual trade-off between efficiency and robustness. (By robustness of an estimator, we mean that it retains consistency properties under a broader range of models than the one upon which it is based.) The consistency of our naive estimating procedure only depends on the first moment specification by the MM model. The results in Propositions 2 and 3 allow us to evaluate the efficiency of our estimators relative to the maximum likelihood estimators when a particular departure from the naive model is specified.

For example, consider the randomized clinical trial situation as above, where the responses are conditionally Poisson and there is a single binary covariate Z indicating treatment assignment, taking values 0 or 1 with equal probability. Suppose also that Z is independent of the follow-up process which starts at time one and ends at random time T . If the baseline cumulative intensity follows a parametric model $\Lambda_0(t) = \exp(\alpha)h_0(t)$, where $h_0(t)$ is some known increasing function of time, such as $h_0(t) = t$ corresponding to a constant intensity rate, then the inference for the treatment effect β now depends on the $\{Y_k\}$ only through the total event counts $Y^+ = \sum_{k=1}^K Y_k$ for each subject. Suppose we postulate a particular

parametric random effects model, namely a Poisson regression model with gamma frailties, i.e.,

$$Y^+|Z, T, \psi \sim \text{Poisson}(\psi h_0(T)e^{\alpha+Z\beta}), \quad \psi \sim \text{Gamma}(\text{Mean}=1, \text{variance}=\kappa), \quad (8)$$

where $h_0(T) \exp(\alpha) = \mathcal{T}$. The Fisher information is given by the 3×3 matrix \ddot{I} , say, where $(\ddot{I})_{qr} = -E\partial_q\partial_r \log p(Y^+|Z, T; \kappa, \alpha, \beta)$, with $q, r = 0, 1, 2$, $\partial_0 = \partial_\kappa$, $\partial_1 = \partial_\alpha$ and $\partial_2 = \partial_\beta$. Such a gamma-mixture of Poisson processes has been discussed by Lawless (1987), and according to his Eqn.(3.9), we have

$$(\ddot{I})_{j0} = 0, \quad (\ddot{I})_{jk} = E\{Z^{j+k-2}\phi_Z(1 + \kappa\phi_Z)^{-1}\}, \quad j, k = 1, 2$$

where $\phi_Z = \mathcal{T}e^{Z\beta}$. Taking the expectation over Z and inverting \ddot{I} , we see that the asymptotic variance of $\hat{\beta}_{ML}$, the MLE of β under the model (8), is given by

$$\text{Avar}(\hat{\beta}_{ML}) = \frac{1}{n}(\ddot{I}^{-1})_{22} = \frac{2}{n} \frac{1}{E\{\mathcal{T}(1 + \kappa\mathcal{T})^{-1}\}} + \frac{2}{n} \frac{1}{E\{\mathcal{T}e^\beta(1 + \kappa\mathcal{T}e^\beta)^{-1}\}}. \quad (9)$$

To simplify the notation in the following discussion, we define $A_z = (2n^{-1})/[E\{\phi_z(1 + \kappa\phi_z)^{-1}\}]$ for $z = 1, 2$, where $\phi_z = \mathcal{T}e^{z\beta}$. Then (9) is simply $\text{Avar}(\hat{\beta}_{ML}) = A_0 + A_1$. For specified parameter values, it is possible to compute (9) numerically or to use simulation to estimate the expectations therein. The asymptotic relative efficiency (*ARE*) is then obtained from the ratio $ARE = \text{Avar}(\hat{\beta}_{ML})/\text{Avar}(\hat{\beta})$, where the denominator is given by (7). However, if follow-up time T and hence $\mathcal{T} = h_0(T) \exp(\alpha)$ is constant across individuals, it can be seen that (7) and (9) coincide, in which case full efficiency $ARE = 1$ is achieved by our estimator $\hat{\beta}$. This suggests that we can perform a Taylor expansion of $\phi_z(1 + \kappa\phi_z)^{-1}$ around the expectation $E\phi_z = \eta_z$, say, for $z = 0, 1$, to obtain a lower bound for (9) and hence a lower bound for the *ARE*. Carrying out the details, a fourth order expansion, leads to the inequalities

$$E\left(\frac{\phi_z}{1 + \kappa\phi_z}\right) \leq \left(\frac{\eta_z}{1 + \kappa\eta_z}\right) \left\{1 - \frac{\eta_z^{-1}\kappa V_2^z}{(1 + \kappa\eta_z)^2} + \frac{\eta_z^{-1}\kappa^2 V_3^z}{(1 + \kappa\eta_z)^3}\right\} \quad \text{for } z = 0, 1,$$

due to the negativeness of the fourth order residual of the Taylor expansion. Here $V_2^z = \text{Var}(\phi_z)$ and $V_3^z = E(\phi_z - \eta_z)^3$. Then, using the inequality $(1 + a)^{-1} \geq 1 - a$ for real a , we obtain

$$A_z = \frac{2}{n} \frac{1}{E\{\phi_z(1 + \kappa\phi_z)^{-1}\}} \geq \frac{2}{n\eta_z} \left\{ (1 + \kappa\eta_z) + \left(\frac{\kappa\eta_z}{1 + \kappa\eta_z}\right) \omega_2^z - \left(\frac{\kappa\eta_z}{1 + \kappa\eta_z}\right)^2 \omega_3^z \right\}$$

for $z = 0, 1$, where $\omega_2^z = V_2^z/\eta_z^2 = \text{Var}(\mathcal{T})/(E\mathcal{T})^2$ and $\omega_3^z = V_3^z/\eta_z^3 = E(\mathcal{T} - E\mathcal{T})^3/(E\mathcal{T})^3$ are the same for $z = 0, 1$. This yields a lower bound for $\text{Avar}(\hat{\beta}_{ML}) = A_0 + A_1$ which, divided

by (7), leads to a lower bound of the ARE as an explicit function of the overdispersion parameter κ . It is of interest to see how the ARE behaves when the overdispersion becomes small or large. The bound obtained above enables us to produce $ARE \geq 1 - O((\kappa ET)^2)$ for small κ , and $ARE \geq (1 + \omega_2^z)^{-1} + O((\kappa ET)^{-1}) = (ET)^2/ET^2 + O((\kappa ET)^{-1})$ for large κ . Note that if the variance of \mathcal{T} is small, the ARE of our procedure is high even if there is large overdispersion. Our small overdispersion result is consistent with the finding by Cox (1983), which considered κ of order $n^{-1/2}$. In all these circumstances, the estimator $\hat{\beta}$ is approximately efficient, and thus, being consistent under a more general model (MM), should be preferred.

6. MEASUREMENT ERROR IN COVARIATES

We now turn to the situation where some components of the covariate vector may be subject to measurement error, that is $Z_k \neq X_k$. In this case the naive estimators (1) may be inconsistent and it is not so simple to solve the bridge equations (3,4). For this section we shall make some simplifying assumptions. First we assume that the covariates do not vary with time, i.e., $X_k = X$ and $Z_k = Z$. Second, we assume that the follow-up process H^* is independent of the covariates Z and X . Now the equations (3,4) become equivalent to:

$$m_k E[e^{Z'b}] = \Lambda_k E[e^{X'\beta}], \quad k = 1, \dots, K \quad (10)$$

$$E[Z e^{Z'b}] / E[e^{Z'b}] = E[Z e^{X'\beta}] / E[e^{X'\beta}] \quad (11)$$

(The equation (11) is obtained from dividing (4) by the sum of (3) over $k = 1, \dots, K$.) If we have a validation data set (X_i, Z_i) , $i = 1, \dots, n'$, we can approximate the expectations above by the sample averages and use methods in Proposition 1 to form adjusted estimates that will be consistent. This approach does not require us to postulate a measurement error model of the relation between X and Z .

Often, some components of the covariate p -vector are measured without error – e.g. treatment assignment (or gender) in the NPC trial data. Let $X = (\tilde{X}', A)'$, $Z = (\tilde{Z}', A)'$, where the sub-vector A is measured without error. Also let the regression parameter $\beta = (\tilde{\beta}', \gamma)'$, and the naive parameter $b = (\tilde{b}', g)'$, corresponding to the partition of Z . In this case from (10,11), we may obtain:

Proposition 4. Suppose A is independent of \tilde{X} , \tilde{Z} , (and of the follow-up process H^*). Then the naive estimator of its corresponding regression coefficient is consistent, i.e., $g = \gamma$.

Proof. We rewrite (11) as

$$\frac{E \left(\begin{array}{c} \tilde{Z} \\ A \end{array} \right) \exp(\tilde{Z}'\tilde{b} + A'g)}{E \exp(\tilde{Z}'\tilde{b} + A'g)} = \frac{E \left(\begin{array}{c} \tilde{Z} \\ A \end{array} \right) \exp(\tilde{X}'\tilde{\beta} + A'\gamma)}{E \exp(\tilde{X}'\tilde{\beta} + A'\gamma)}$$

Using the independence of A and (\tilde{Z}, \tilde{X}) , the A -coordinates in this system of equations yield:

$$\frac{E[A \exp(A'g)]}{E[\exp(A'g)]} = \frac{E[A \exp(A'\gamma)]}{E[\exp(A'\gamma)]} \quad (12)$$

Hence clearly $g = \gamma$ is a solution. To show it is unique, we note that (12) can be written as: $\partial_g f(g) = \partial_\gamma f(\gamma)$, where $f(g) = \log E[\exp(A'g)]$. Note that

$$\partial_g \partial'_g f(g) = \frac{E[AA' \exp(A'g)]}{E[\exp(A'g)]} - \frac{\{E[A \exp(A'g)]\}^{\otimes 2}}{\{E[\exp(A'g)]\}^2} = \tilde{E}AA' - (\tilde{E}A)^{\otimes 2}$$

which is positive definite except in the trivial case when $a'A$ is constant with probability one for some non-zero a . (Here $\tilde{E}(v)$ denotes $E\left[v \left(\frac{\exp(A'g)}{E[\exp(A'g)]}\right)\right]$ for any v .) Thus $f(g)$ is strictly convex and $g = \gamma$ is the unique solution to (12).

For example, Proposition 4 applies when A is a treatment assignment variable in a randomized clinical trial, and is, by design, independent of the covariates and the followup process. Proposition 4 implies that A 's effect is consistently estimated using the partial likelihood estimator (1), which neglects measurement error and random effects.

In order to proceed further and obtain analytic solutions to (10,11), we will need to make modeling assumptions on the form of the measurement error, specifically the joint distribution of X and Z . We say that the measurement error structure follows a *conditional normal* (CN) model if we have

$$\tilde{X} | Z \sim N(C'_0 + C'Z, \Sigma) \quad (13)$$

for some general vector C_0 , matrix C and covariance matrix Σ . We partition the matrix $C' = (\Omega', C'_A)$ so that $C'Z = \Omega'\tilde{Z} + C'_A A$. The model (13) holds for example if X and Z are jointly normal, or (13) may arise from a regression calibration model (Carroll et al. 1995, Secs. 1.3, 3). For this model we can solve (10,11):

Proposition 5. If the CN model holds then:

$$g = \gamma + C_A \tilde{\beta}, \quad \tilde{b} = \Omega \tilde{\beta}, \quad m_k = \Lambda_k \exp\{C_0 \tilde{\beta} + \frac{1}{2} \tilde{\beta}' \Sigma \tilde{\beta}\}. \quad (14)$$

Proof. We can solve the equations (10) and (11) by first conditioning on Z when taking the expectations on the right hand sides. Using the expression for the moment generating function of the multivariate normal distribution of \tilde{X} given Z we have that

$$E(e^{X'\beta} | Z) = E(e^{\tilde{X}'\tilde{\beta} + A'\gamma} | Z) = \exp\{(C_0 \tilde{\beta} + \frac{1}{2} \tilde{\beta}' \Sigma \tilde{\beta}) + \tilde{Z}'(\Omega \tilde{\beta}) + A'(\gamma + C_A \tilde{\beta})\}.$$

A solution (14) can now be easily read off as given by the proposition. The uniqueness of this solution can be demonstrated in a similar way as in the proof of Proposition 4.

We see immediately that, when measurement error is present, the naive estimators are no longer consistent for the original parameters. We invert the relations (14) to obtain

$$\gamma = g - C_A \Omega^{-1} \tilde{b}, \quad \tilde{\beta} = \Omega^{-1} \tilde{b}, \quad \Lambda_k = m_k \exp\{-C_0 \Omega^{-1} \tilde{b} - \frac{1}{2} \tilde{b}' \Pi \tilde{b}\}, \quad (15)$$

where $\Pi \equiv (\Omega^{-1})' \Sigma \Omega^{-1}$. The third equations in (14) and (15) lead to the following relations for the baseline cumulative intensity $\Lambda_0(t) = \sum_{k \leq t} \Lambda_k$:

$$m_0(t) = \Lambda_0(t) \exp\{C_0 \tilde{\beta} + \frac{1}{2} \tilde{\beta}' \Sigma \tilde{\beta}\} \text{ and } \Lambda_0(t) = m_0(t) \exp\{-C_0 \Omega^{-1} \tilde{b} - \frac{1}{2} \tilde{b}' \Pi \tilde{b}\}. \quad (16)$$

The equations (15) and (16) tell us how to form consistent estimators for the parameters of interest from the naive estimators in (1). However the constants C_A , Ω and Π relating to the measurement error model may not be known and in general will need to be estimated from a separate “validation” study, as is commonly the case when there is measurement error (e.g., Carroll et al. 1995, p.12). For the NPC trial data, such estimates were available — see the next section. In many situations we can simply take the constant vector C_0 of the CN model to be zero by suitably re-centering the covariates. For example, $C_0 = 0$ if \tilde{X} and \tilde{Z} are both re-centered to have zero mean for subjects with covariate A equal to zero. From now on we will take $C_0 = 0$.

The CN model also arises from a normal additive (NADD) model, where it is assumed

$$\tilde{Z} = \tilde{X} + U, \quad \tilde{X} \sim N(0, \Sigma_{\tilde{X}}), \quad U \sim N(0, \Sigma_U), \quad (17)$$

and \tilde{X} and U are mutually independent and both independent of A .

It can be seen that the NADD model gives a CN model in which $C_0 = C_A = 0$, and

$$\Omega = \Sigma_{\tilde{Z}}^{-1} \Sigma_{\tilde{X}}, \quad \Sigma = \Sigma_U \Sigma_{\tilde{Z}}^{-1} \Sigma_{\tilde{X}}, \quad \text{where } \Sigma_{\tilde{Z}} = \Sigma_{\tilde{X}} + \Sigma_U.$$

The matrix Ω is called the “attenuation” matrix. When \tilde{X} is a scalar, we see that the NADD model implies that $0 < \Omega < 1$.

Now we consider the asymptotic variances of these estimators. Similar to Section 5, we consider a special situation when the $\{Y_{ik}\}$ are conditionally Poisson distributed, with constant overdispersion parameter $\kappa = \text{Var}(\psi | Z^*, X^*, H^*)$. Suppose we have a conditional normal (CN) model for the measurement error structure. Consider the naive estimators \hat{b} and \hat{m}_k 's first. Naive and robust calculation of the asymptotic variance of the naive MLE \hat{b} are based on $(n\mathcal{I})^{-1}$ and $n^{-1}\mathcal{I}^{-1}\mathcal{V}\mathcal{I}^{-1}$, respectively. An argument similar to that in Section 5 yields:

$$\mathcal{V} - \mathcal{I} = E \left[\left\{ \kappa \mathcal{T}^2 E(e^{2X'\beta} | Z) + \mathcal{T}^2 \text{Var}(e^{X'\beta} | Z) \right\} \left(Z - \frac{EZ e^{Z'b}}{E e^{Z'b}} \right)^{\otimes 2} \right] \quad (18)$$

which is positive definite. Hence $Avar(\hat{b}_j) \geq Avar_N(\hat{b}_j)$, for all $j = 1, \dots, p$. Similarly it can be shown that $Avar(\hat{m}_0(t)) \geq Avar_N(\hat{m}_0(t))$, for all $t = 1, \dots, K$. The inequalities are strict if $\kappa > 0$ and there exists no non-zero constant vector $a = (a_1, \dots, a_p)$ such that $Var(\sum_{q=1}^p a_q Z_q) = 0$. We note that the variance inflation in (18) contains a term proportional to κ due to random effects, as well as another proportional to $Var(e^{X'\beta}|Z)$ contributed by the measurement error in Z .

Because, unlike the situation in Section 4, the naive MLEs are not consistent for (Λ_k, β) , we must use the double-sandwich formula to estimate the asymptotic variances of the adjusted estimates $\hat{\Lambda}_k$ and $\hat{\beta}$ in general. However when A is the treatment variable in a randomized clinical trial, in which \tilde{X} is assumed independent of A conditional on \tilde{Z} , we have $C_A = 0$ and $\gamma = g$. In this situation, the asymptotic variances of particular interest are $Avar(\hat{\gamma}) = Avar(\hat{g})$ and $Avar(\hat{\beta}) = \Omega^{-1}Avar(\hat{b})(\Omega^{-1})'$. (These can be obtained either by directly using the delta method or formally calculating the D -matrix in Section 3). The asymptotic variance of $\hat{\Lambda}_0(t)$ can also be obtained from (16) as

$$\begin{aligned} Avar\{\hat{\Lambda}_0(t)\} &= e^{-\tilde{b}'\Pi\tilde{b}}[Avar\{\hat{m}_0(t)\} + m_0^2(t)\tilde{b}'\Pi'\{Avar(\hat{b})\}\Pi\tilde{b} \\ &\quad - m_0(t)Acov\{\hat{m}_0(t), \hat{b}\}\Pi\tilde{b} - m_0(t)\tilde{b}'\Pi'Acov\{\hat{b}, \hat{m}_0(t)\}]. \end{aligned} \quad (19)$$

(Here we have taken C_0 to be zero in the CN model.) In the above expressions, the asymptotic variances and covariances for the naive estimates are given by equation (20) in Appendix A, and \tilde{b} and $m_0(t)$ can be estimated by the naive estimates, \hat{b} and $\hat{m}_0(t)$, respectively.

To illustrate the meaning of these results consider the case when $b = (\tilde{b}, g)$ consists of two scalars (i.e., $p = 2$) and we have $C_A = 0$ in the CN model (e.g., as in the NADD model). Then $\gamma = g$, $\tilde{\beta} = \Omega^{-1}\tilde{b}$, $Avar(\hat{b}) > Avar_N(\hat{b})$, and $Avar(\hat{g}) > Avar_N(\hat{g})$. However, for inference on γ , asymptotically, the magnitude of the adjusted \mathcal{Z} value

$$|\mathcal{Z}_{Adj.}| = \frac{|\hat{\gamma}|}{\sqrt{Avar(\hat{\gamma})}} = \frac{|\hat{g}|}{\sqrt{Avar(\hat{g})}} < \frac{|\hat{g}|}{\sqrt{Avar_N(\hat{g})}} = |\mathcal{Z}_N|$$

Thus when adjusted, the significance of the regression coefficient γ is diminished. For inference on $\tilde{\beta}$, the regression coefficient for the predictor measured with error, asymptotically

$$|\mathcal{Z}_{Adj.}| = \frac{|\hat{\tilde{\beta}}|}{\sqrt{Avar(\hat{\tilde{\beta}})}} = \frac{|\Omega^{-1}\hat{\tilde{b}}|}{\sqrt{Avar(\Omega^{-1}\hat{\tilde{b}})}} = \frac{|\hat{\tilde{b}}|}{\sqrt{Avar(\hat{\tilde{b}})}} < \frac{|\hat{\tilde{b}}|}{\sqrt{Avar_N(\hat{\tilde{b}})}} = |\mathcal{Z}_N|.$$

Again, the significance of the regression coefficient is diminished, even though the adjusted estimate $\hat{\tilde{\beta}}$ is greater in magnitude than the naive one, because the adjusted asymptotic variance more than compensates for this. Of course these results pertain to asymptotic limits — in finite samples, these relations may not hold exactly.

In the above discussion we have been treating the parameters Ω^{-1} and Π as known. In practice, as noted in Section 3, they will usually need to be estimated from some auxiliary or “validation” data set and will be themselves subject to sampling error. In the example of the next section we show how uncertainty in these parameters can be incorporated into inferences on γ , $\tilde{\beta}$ and $\Lambda_0(t)$.

In summary, the usual partial likelihood procedure will give a consistent estimate for the treatment effect γ . However, in general the naive asymptotic variance from $\hat{\mathcal{I}}^{-1}$ does not account for the model complications such as random effects and measurement errors neglected by the naive model, and the robust variance should be used instead, which is valid for evaluating the asymptotic variance of $\hat{\gamma}$ whatever the true probability model is. For the overdispersed Poisson models, the naive asymptotic variance tends to be an underestimate. When the robust asymptotic variance is used instead, a smaller \mathcal{Z} -value for the Wald statistic will usually result. Naive estimates of regression coefficients for variables measured with error are attenuated. All these features are exhibited in the application in Section 7.

7. APPLICATION TO NPC TRIAL DATA

We now illustrate these methods with an analysis of the SCC recurrences in the NPC trial reported in Clark et al. (1996) and described in Section 1. We use two covariates, namely treatment assignment indicator ($A=1$ for Se supplemented group, $A=0$ for placebo) and \tilde{Z} , the logarithm of baseline plasma Se status, centered to have zero mean for placebo subjects. Se status is measured in units of ng/ml. Because of imprecise measurement instrumentation and natural temporal biological variability, \tilde{Z} is measured with error, with its hypothetical true value \tilde{X} unknown. We use the simple normal additive model (NADD), whereby $\tilde{Z} = \tilde{X} + U$, as introduced in the previous section. For the parameters of this model we use values $\Sigma_{\tilde{X}} = 0.106^2$ (0.048^2), $\Sigma_U = 0.151^2$ (0.021^2), being scalars since \tilde{Z} is one-dimensional. These values, with standard errors (*s.e.*'s) in parentheses, were estimated from a validation data set based on replicate plasma Se measurements in placebo patients. The details are given in Appendix B, where we also show how the same data set can be used to check the adequacy of the NADD model assumptions. A similar strategy was used in Turnbull et al. (1997, Sec. 6.3), but their estimates differ from ours because they used an earlier and smaller interim data set. From these values, we obtain the attenuation factor of $\Omega^{-1} = (\Sigma_{\tilde{X}} + \Sigma_U)/\Sigma_{\tilde{X}} = 3.01$ (0.46), and $\Pi = \Omega^{-1}\Sigma\Omega^{-1} = \Omega^{-1}\Sigma_U = 0.261^2$ (0.108^2) where $\Sigma = \Sigma_{\tilde{X}}\Sigma_U/(\Sigma_{\tilde{X}} + \Sigma_U)$. The standard errors given for Ω^{-1} and Π were obtained from those of Σ_X and Σ_U through use of the delta method or “propagation of errors” formulae — details are also shown in Appendix B.

Table 1 shows the results of fitting the semiparametric models. In Models 1a,1b, we fit only the treatment assignment variable and see that, with a naive analysis using a variance

estimator based on $n^{-1}\mathcal{I}^{-1}$ which neglects any potential random effects (Model 1a), there appears to be a significant harmful effect of treatment. When robust variance estimates are used which automatically account for any random effects (Model 1b), the magnitude of the effect is unchanged, but the effect is no longer statistically significant. This agrees qualitatively with the results of a logrank test based on time to first SCC occurrence only, that was reported by Clark et al. (1996). In Models 1c,1d, we also adjust for the presence of the covariate \tilde{Z} , baseline plasma Se status. Again the treatment effect becomes insignificant once random effects and measurement errors are taken into account. The significance of the baseline Se status is diminished but Se status still remains highly significant as a prognostic factor for SCC occurrence, thus agreeing with earlier epidemiologic evidence. It should be noted that these results are based on regarding the measurement error model (or “validation”) parameters $\Omega^{-1} = 3.01$, and $\Pi = 0.261^2$ as known. Sensitivity analyses can be based on replacing $\Omega^{-1} = 3.01$ by $\Omega^{-1} \pm s.e.(\Omega^{-1})$, say, but the adjusted \mathcal{Z} values will remain unchanged for any value of Ω^{-1} , since the latter contributes to the adjusted estimates of the regression parameters and their standard errors in the same proportion. However, for confidence intervals, we need to account for the extra variation stemming from the fact the measurement error model parameters are only estimated. Because $\hat{\beta} = \Omega^{-1}\hat{b}$, an upper bound for the s.e. of $(\hat{\beta})$ can be found using results on propagation of errors for products — e.g., see Taylor (1997, Eqn 3.8). Using the robust s.e. of 0.320 for \hat{b} and estimates given above and in Table 1, we have

$$\begin{aligned} s.e.(\hat{\beta}) &\leq |\hat{\beta}| \left\{ \frac{s.e.(\hat{b})}{|\hat{b}|} + \frac{s.e.(\Omega^{-1})}{\Omega^{-1}} \right\} \\ &= 2.181 \left\{ \frac{0.320}{0.690} + \frac{0.46}{3.01} \right\} = 1.281 \end{aligned}$$

This standard error for $(\hat{\beta})$ can be used in place of the one entered in the table (0.963) which ignored the uncertainty in Ω . A nominal 95% confidence interval constructed using this upper bound of $s.e.(\hat{\beta})$ is conservative and contains zero. The s.e. of $\hat{\gamma}$ is unaffected since it does not depend on the validation parameters under the NADD model, or more generally due to Proposition 4.

[Table 1 and Figure 1 about here.]

Figure 1 shows the naive and adjusted estimates of the baseline cumulative intensity function. The naive estimate $\hat{m}_0(t)$ is larger than the adjusted consistent estimator $\hat{\Lambda}_0(t)$ as stated in Section 6, although the difference is extremely small and hard to distinguish. This is because the ratio of the two estimators, given by $e^{\frac{1}{2}(\hat{b})'(\Omega^{-1})'\Sigma\Omega^{-1}\hat{b}} = 1.0164$, is very close

to one. In general, when this happens, the influence of covariate measurement error on the baseline cumulative intensity estimator is negligible. In Figure 1 we have also plotted the pointwise 95% confidence bands for $\hat{m}_0(t)$ based on the naive asymptotic variance obtained from (22) based on $(nI)^{-1}$, as well as bands based on the robust asymptotic variance (19) for the consistent estimator $\hat{\Lambda}_0(t)$. In this case, when account is taken of the fact that the measurement error model parameter Π is not known but estimated from the validation data set, the increase in the standard error is very small — the increase in the coefficient of variation of $\hat{\Lambda}_0(t)$ is less than 0.3%. (This comes from a propagation of errors formula similar to that used for $s.e.(\hat{\beta})$.)

A parametric model of constant baseline intensity rate with $\Lambda_0(t) = t \exp(\alpha)$ is suggested by Figure 1. Upon fitting such a parametric model, the results, corresponding to those for Models 1a-1d in Table 1, are shown as Models 2a-2d in Table 1. Results from Models 2c and 2d, based on a Poisson process regression model, can be compared with results from a negative binomial regression model used in Tables I and II of Turnbull et al. (1997). However, it should be recognized that Turnbull et al. (1997) used an earlier and much smaller interim data set from the NPC trial. One benefit of fitting a semiparametric model is to validate a fully parametric model, such as here using Figure 1 to justify the adequacy of a constant baseline intensity rate; parametric models have advantages for prediction, interpretation and communication. It is also noted that the results of $s.e.(\hat{\beta})$ in Model 2d neglect to account for the extra variation stemming from the fact the measurement error model parameters Ω^{-1} and Π are only estimated, and considerations similar to the semi-parametric situation (Model 1d) apply. In this case, incorporating uncertainty in these estimates leads to an upper bound for $s.e.(\hat{\beta})$ of 1.297 instead of 0.963.

8. DISCUSSION OF ASSUMPTIONS

It should be noted that, whereas inferences concerning treatment effect, as in our NPC example, are robust, those which involve covariates measured with error are model dependent. In particular, the NADD model assumption can be checked if there is a validation data set in which pairs $(\tilde{X}_i, \tilde{Z}_i)$ are available. In that case, normal probability plots of the components of the $\{\tilde{X}_i\}$ and $\{\tilde{Z}_i\}$ can reveal any departures from normality, and scatter plots of $\{\tilde{Z}_i - \tilde{X}_i, \tilde{X}_i\}$ or other tests can be used to check independence of $U = \tilde{Z} - \tilde{X}$ and \tilde{X} — see Appendix B.

Many of our results can be easily adapted to an arbitrary link function, not necessarily the loglinear one. However, the Propositions 4 and 5 do depend critically on the loglinear link function that relates the covariates to the response rates. As mentioned in Section 2,

a log-linear link function is the natural one and often reasonable, possibly after suitable transformation of the covariates (see, e.g., Andersen et al. 1993, page 547). Of course, for the two-sample comparison, use of the loglinear link is not restrictive because Z is then only a zero-one indicator. It is possible to perform diagnostic checks for the log-linearity. For example, all subjects can be binned into L sets, say, according to covariate value Z . When the crude event rate for each set is plotted on the log scale versus mean values of each covariate component for that set, the resulting p scatter plots should appear approximately linear. In our NPC example, we binned into sets of approximately 40 subjects according to baseline plasma Se status and treatment group. Upon fitting a loess line through a scatter plot, not shown, of log SCC event rates versus log baseline plasma Se status, the result was approximately linear, suggesting adequacy of the loglinear link function.

A fundamental assumption implied by the conditions stated in Section 2, is that the underlying event rates should be independent of the observation process H^* . In the case of right censoring only, this would be satisfied if the length of followup or observation time for each subject is independent of the event process of that subject. However certain situations are excluded, such as for example when: (i) a subject is more likely to leave the study earlier if he has a higher frequency of events; or (ii) a subject is withdrawn from the study as soon as he has experienced a fixed number, r say, of events. In such a case, subjects with higher frailties (higher event rates) would be less likely to be still at risk at later time periods, resulting in an underestimate of the intensity function there.

Lawless and Nadeau (1995, p.164) propose several tests of the independent observation time assumption. For our NPC trial data, we performed their two Wald-type tests which are based on including an extra covariate of either (a) the length of observation time or follow-up time, T say; or (b) an indicator on whether T is longer than the median (here 2795 days). The second test (b) yielded a Z -value of -0.65 for the new covariate, which is not significant. This indicates no evidence against the independence assumption. On the other hand, the first test (a) did indicate a marginally significant negative relation between the event frequency and the follow-up length. However as Lawless and Nadeau (1995) point out, this test is highly sensitive to influential observations. We calculated the correlation between the event rate (number of events divided by T) and the follow-up length (T). This was done separately for the placebo and Se group patients. The correlations are -0.0194 and -0.0931 respectively, whereas the 3% trimmed correlations were $+0.0039$ and $+0.0032$. Thus the data seemed generally consistent with the independence assumption. Also there was nothing in the protocol of the design or conduct of the followup that would lead to a suspicion that there should be nonindependence of the event processes and the observation times.

Finally, we remark that the methodology here is based on large sample asymptotic theory. Although the data set we used in our illustration was quite large, it would be of interest to investigate finite sample properties of the procedure.

APPENDIX A: ASYMPTOTIC VARIANCE ESTIMATORS

In this Appendix, we give expressions for robust variance estimates for the naive estimators (1). These robust expressions are valid in presence of random effects and measurement error. Variances of the naive estimates involve quantities V and I via result (A) of Proposition 1. The robust variance estimators are constructed from \hat{I} and \hat{V} , obtained by substituting the expectations in I and V with sample analogs and replacing the naive parameters b and m by the naive estimates \hat{b} and \hat{m} , respectively.

The parameter (row)-vector s' can be partitioned into a K -dimensional sub-vector $m' = (m_1, \dots, m_K)$ and a p -dimensional sub-vector b . We adopt the fairly standard notation

$$S_k^{(0)} = \sum_i H_{ik} e^{Z'_{ik} \hat{b}}, \quad S_k^{(1)} = \sum_i H_{ik} Z_{ik} e^{Z'_{ik} \hat{b}}, \quad S_k^{(2)} = \sum_i H_{ik} Z_{ik} Z'_{ik} e^{Z'_{ik} \hat{b}}, \quad E_k = S_k^{(1)} / S_k^{(0)}.$$

$$\hat{\mathcal{I}} = -\partial_b \partial_b' \mathcal{L} = \sum_k \left(\sum_i H_{ik} Y_{ik} \right) (S_k^{(2)} / S_k^{(0)} - E_k E_k').$$

$\hat{\mathcal{I}}$ is exactly the same as the sample information matrix obtained by taking second order derivative of the partial log likelihood \mathcal{L} with respect to b . Thus $\hat{\mathcal{I}}^{-1} = \widehat{Avar}_N(\hat{b})$ is the asymptotic variance matrix used in the naive partial likelihood analysis that ignores any random effects or measurement error, which is not robust.

Define the K -dimensional sub-vectors $\{A_{ik}\}$ by components:

$$(A_{ik})_l = \delta_{lk} \left(\sum_j H_{jl} e^{Z'_{jl} \hat{b}} \right)^{-1} - \hat{m}_l E_l' \hat{\mathcal{I}}^{-1} (Z_{ik} - E_k), \quad l = 1, \dots, K$$

and the p -dimensional sub-vector $(B_{ik}) = \hat{\mathcal{I}}^{-1} (Z_{ik} - E_k)$. Here δ_{lk} is the Kronecker delta.

By the sandwich formula, we estimate the robust asymptotic variance by $\widehat{Avar}(\hat{s}) = (n\hat{I})^{-1} (n\hat{V}) (n\hat{I})^{-1}$, where $\hat{s}' = (\hat{m}', \hat{b}')$. The result is

$$\widehat{Avar} \begin{bmatrix} \hat{m} \\ \hat{b} \end{bmatrix} = \sum_i \left\{ \sum_k H_{ik} (Y_{ik} - \hat{m}_k e^{Z'_{ik} \hat{b}}) \begin{bmatrix} A_{ik} \\ B_{ik} \end{bmatrix} \right\}^{\otimes 2}.$$

These equations enable us to calculate the variance estimators for \hat{m}_k and \hat{b} . We can show that the robust asymptotic variance estimator for \hat{b} can be put in the sandwich form

$$\widehat{Avar}(\hat{b}) = \hat{\mathcal{I}}^{-1} \hat{V} \hat{\mathcal{I}}^{-1} \text{ where } \hat{V} = \sum_i \left\{ \sum_k H_{ik} (Y_{ik} - \hat{m}_k e^{Z'_{ik} \hat{b}}) (Z_{ik} - E_k) \right\}^{\otimes 2}.$$

Here \mathcal{V} is the sample variance estimator for the score function corresponding to the partial log likelihood \mathcal{L} .

The asymptotic variance for $\hat{m}_0(t)$, the naive estimate of the baseline cumulative intensity, can then be expressed in terms of those of \hat{m}_k 's. The asymptotic covariance matrix estimator of the vector $(\hat{m}_0(t), \hat{b})'$ and $(\hat{m}_0(s), \hat{b})'$ is given by the following formula:

$$\widehat{Acov} \left\{ \begin{bmatrix} \hat{m}_0(t) \\ \hat{b} \end{bmatrix}, \begin{bmatrix} \hat{m}_0(s) \\ \hat{b} \end{bmatrix} \right\} \equiv \begin{bmatrix} \widehat{Acov}\{\hat{m}_0(t), \hat{m}_0(s)\} & \widehat{Acov}\{\hat{m}_0(t), \hat{b}\} \\ \widehat{Acov}\{\hat{b}, \hat{m}_0(s)\} & \widehat{Avar}(\hat{b}) \end{bmatrix} = \sum_i G_i(t)G_i(s)' \quad (20)$$

$$\text{where } G_i(t) \equiv \sum_k H_{ik}(Y_{ik} - \hat{m}_k e^{Z'_{ik}\hat{b}}) \begin{bmatrix} \sum_{l \leq t} (A_{ik})_l \\ B_{ik} \end{bmatrix}.$$

In particular,

$$\widehat{Avar} \{ \hat{m}_0(t) \} = \sum_i \left[\sum_{l \leq t} \sum_k H_{ik} (Y_{ik} - \hat{m}_k e^{Z'_{ik}\hat{b}}) \left\{ \delta_{lk} (\sum_j H_{jl} e^{Z'_{jl}\hat{b}})^{-1} - \hat{m}_l E'_l \hat{\mathcal{I}}^{-1} (Z_{ik} - E_k) \right\} \right]^{\otimes 2}. \quad (21)$$

On the other hand, the naive asymptotic variance can be obtained by

$$\widehat{Avar}_N \{ \hat{m}_0(t) \} = \sum_{l \leq t} \sum_{k \leq t} Acov_N(\hat{m}_l, \hat{m}_k) = \sum_{k \leq t} \hat{m}_k (\sum_i H_{ik} e^{Z'_{ik}\hat{b}})^{-1} + \sum_{l \leq t} \sum_{k \leq t} \hat{m}_l E'_l \hat{\mathcal{I}}^{-1} E_k \hat{m}_k. \quad (22)$$

In the continuous limit, (22) is consistent with the formula on page 505 of Andersen et al. (1993). When random effects or measurement errors exist, however, we should use the robust variance estimator in (21).

APPENDIX B: VALIDATION DATA SET

In Section 6, it was assumed that the parameters of the measurement error model were known. Usually, however, this will not be the case and they must be estimated from an auxiliary or “validation” data set. In this Appendix we describe the estimation of the parameters in the NADD model used for the analysis of the NPC trial data set described in Section 7.

For all patients in the NPC trial, plasma Se measurements were taken serially at approximate six month intervals, and not just at baseline (randomization). Assuming stationarity, the repeated readings from each placebo patient should represent replicated measurements of \tilde{X} for that patient. We are including the natural biological temporal variation in Se plasma levels as well as instrument error as components of the “measurement” error. Recall that \tilde{X} represents the long run mean level for an individual untreated patient. Of course, we do not include Se readings of treated patients since their subsequent levels would be affected

by the nutritional supplements of Se they were taking. The stationarity assumption can be examined by checking that the group mean Se levels remain approximately constant over time and by using control chart techniques (Montgomery 1997) on the longitudinal series of measurements in samples of individual placebo subjects. For the NPC trial data, the stationarity assumption seemed reasonable. There were 647 placebo patients with baseline Se levels recorded. Let $\tilde{Z}_{i1}, \tilde{Z}_{i2}, \dots, \tilde{Z}_{ir_i}$ denote the replicate $\log(\text{Se})$ readings for the i th placebo patient ($i = 1, \dots, 647$). We obtain estimates of $\mu_{\tilde{X}} = \mu_{\tilde{Z}}$ and of $\Sigma_{\tilde{Z}} = \Sigma_{\tilde{X}} + \Sigma_U$ from the baseline readings $\{\tilde{Z}_{i1}\}$:

$$\begin{aligned}\hat{\mu}_{\tilde{X}} = \hat{\mu}_{\tilde{Z}} &= \frac{1}{647} \sum_{i=1}^{647} \tilde{Z}_{i1} &= 4.719 \\ \hat{\Sigma}_{\tilde{Z}} &= \frac{1}{646} \sum_{i=1}^{647} (\tilde{Z}_{i1} - \hat{\mu}_{\tilde{Z}})^2 &= 0.185^2\end{aligned}$$

An estimate of Σ_U can be obtained from the pooled within placebo subject variability:

$$\hat{\Sigma}_U = \frac{\sum_{i=1}^{647} \sum_{j=1}^{r_i} (\tilde{Z}_{ij} - \bar{\tilde{Z}}_i)^2}{\sum_{i=1}^{647} (r_i - 1)} = 0.151^2.$$

Standard errors for $\hat{\Sigma}_{\tilde{Z}}$ and $\hat{\Sigma}_U$ can be based on the variance of the chi-squared distribution: $s.e.(\hat{\Sigma}) \simeq \hat{\Sigma} \sqrt{2/d.f.}$. For $\hat{\Sigma}_{\tilde{Z}}$, we have $d.f. = 647 - 1 = 646$ and the standard error is $s.e.(\hat{\Sigma}_{\tilde{Z}}) = 0.044^2$. For $\hat{\Sigma}_U$, we have $d.f. = \sum_{i=1}^{647} (r_i - 1) = 5225$ and the standard error is $s.e.(\hat{\Sigma}_U) = 0.021^2$. Finally we have $\hat{\Sigma}_{\tilde{X}} = \hat{\Sigma}_{\tilde{Z}} - \hat{\Sigma}_U = 0.106^2$ with standard error bounded above by $s.e.(\hat{\Sigma}_{\tilde{Z}}) + s.e.(\hat{\Sigma}_U) = 0.048^2$ (e.g., Taylor, 1997, Eqn. 3.4).

These estimates in turn lead us to estimates of the attenuation factor:

$$\hat{\Omega}^{-1} = 1 + [\hat{\Sigma}_U / \hat{\Sigma}_{\tilde{X}}] = 3.01 \quad (0.46)$$

and

$$\hat{\Pi} = \hat{\Omega}^{-1} \hat{\Sigma}_U = 0.261^2 \quad (0.108^2).$$

The standard errors in parentheses were calculated from those of $\hat{\Sigma}_U$ and $\hat{\Sigma}_{\tilde{X}}$ by using bounds obtained from standard propagation of errors formulae for products and quotients (e.g., Taylor, 1997, Eqn. 3.8):

$$\begin{aligned}s.e.(\hat{\Omega}^{-1}) = s.e.(\hat{\Sigma}_U / \hat{\Sigma}_{\tilde{X}}) &\leq \left(\frac{\hat{\Sigma}_U}{\hat{\Sigma}_{\tilde{X}}} \right) \left(\frac{s.e.(\hat{\Sigma}_U)}{\hat{\Sigma}_U} + \frac{s.e.(\hat{\Sigma}_{\tilde{X}})}{\hat{\Sigma}_{\tilde{X}}} \right) \\ &= \left(\frac{.151^2}{.106^2} \right) \left(\frac{.021^2}{.151^2} + \frac{.048^2}{.106^2} \right) = 0.455.\end{aligned}$$

Similarly,

$$\begin{aligned}s.e.(\hat{\Pi}) &\leq \hat{\Pi} \left(\frac{s.e.(\hat{\Omega}^{-1})}{\hat{\Omega}^{-1}} + \frac{s.e.(\hat{\Sigma}_U)}{\hat{\Sigma}_U} \right) \\ &= 0.261^2 \left(\frac{.455}{3.01} + \frac{.021^2}{.151^2} \right) = 0.108^2\end{aligned}$$

The validation data set can also be used to check the distributional assumptions on \tilde{X} and U in the normal additive covariate error model (NADD, Section 6). To do this, we restricted ourselves to the 322 placebo patients for whom ten or more serial Se readings were available at the time of the analysis, i.e., those patients with $r_i \geq 10$. For such patients, the mean $\log(\text{Se})$ reading should be a reasonably accurate estimate of the “true” \tilde{X} -value. Thus we replace \tilde{X}_i with $\hat{\tilde{X}}_i = \overline{\tilde{Z}}_i = \frac{1}{r_i} \sum_j \tilde{Z}_{ij}$. The $\{\tilde{Z}_i\}$ are of course the initial $\log(\text{Se})$ readings taken at randomization, that is $\tilde{Z}_i = \tilde{Z}_{i1}$. We can then take U_i to be $\hat{U}_i = \tilde{Z}_i - \hat{\tilde{X}}_i$, ($1 \leq i \leq 322$). A small correlation between the $\{\hat{U}_i\}$ and the $\{\hat{\tilde{X}}_i\}$ would indicate the appropriateness of the independence assumption. Here, this correlation was computed to be -0.087 (A robust 10% trimmed estimate was -0.024). Similarly histograms and probability plots of the $\{\hat{U}_i\}$ and the $\{\hat{\tilde{X}}_i\}$ indicate the adequacy of the normality assumption, see Figure 2. Note that these plots are made before the zero-mean re-centering of the $\log(\text{Se})$ readings.

[Figure 2 about here.]

REFERENCES

- Abu-Libdeh, H., Turnbull, B. W., and Clark, L. C. (1990), “Analysis of Multi-type Recurrent Events in Longitudinal Studies; Application to a Skin Cancer Prevention Trial,” *Biometrics*, 46, 1017-1034.
- Allison, P.D. (1984), *Event History Analysis: Regression for Longitudinal Data*, Sage Publications, Beverly Hills, California.
- Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993), *Statistical Models Based on Counting Processes*, New York: Springer-Verlag,
- Andersen, P. K. and Gill, R. D. (1982), “Cox’s Regression Model for Counting Processes: A Large Sample Study,” *The Annals of Statistics*, 10, 1100-1120.
- Ascher, H. and Feingold, H. (1984). *Repairable Systems Reliability: Modeling, Inference, Misconceptions and Their Causes*, Marcel Dekker, New York.
- Buzas, J. S. (1998), “Unbiased scores in proportional hazards regression with covariate measurement error” *Journal of Statistical Planning and Inference*, 67, 247-257.
- Carroll, R. J., Ruppert, D. and Stefanski, L. A. (1995), *Measurement Error in Nonlinear Models*, Chapman and Hall, New York.
- Clark, L.C., Combs, G. F., Turnbull, B.W., Slate, E.H., Chalker, D.K., Chow, J., Davis, L.S., Glover, R.A., Graham, G.F., Gross, E.G., Krongrad, A., Leshner, J.L., Park, H.K., Sanders, B.B., Smith, C.L., Taylor, J.R. and the Nutritional Prevention of Cancer Study Group. (1996). “Effects of Selenium Supplementation for Cancer Prevention in Patients with Carcinoma of the Skin: A Randomized Clinical Trial.” *Journal of the American Medical Association*, 276 (24), 1957-1963.

- Cook, R. (1995), "The Design and Analysis of Randomized Trials with Recurrent Events," *Statistics in Medicine*, 14, 2081-2098.
- Cook, R. J., Lawless, J. F. and Nadeau, C. (1996) "Robust Tests for Treatment Comparisons Based on Recurrent Event Responses," *Biometrics*, 52, 557-571.
- Cox, D. R. (1972), "Regression Models and Life-tables," *Journal of the Royal Statistical Society B*, 34, 187-207.
- Cox D. R. (1983), "Some remarks on overdispersion," *Biometrika*, 70, 269-274.
- Gail, M. H., Wieand, S. and Piantadosi, S. (1984), "Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates," *Biometrika*, 71, 431-444.
- Huber, P.J. (1967), "The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions," *Proceedings of the 5th Berkeley Symposium on Probability and Statistics 1*, University of California Press, 221-233.
- Jiang, W. (1996), *Aspects of Misspecification in Statistical Models: Applications to Latent Variables, Measurement Error, Random Effects, Omitted Covariates and Incomplete Data*, Ph. D. Thesis, Cornell University.
- Keiding, N., Andersen, P. K. and Klein, J. P. (1997), "The Role of Frailty Models and Accelerated Failure Time Models in Describing Heterogeneity Due to Omitted Covariates," *Statistics in Medicine*, 16, 215-224.
- Lagakos, S. W. and Shoenfeld, D. A. (1984), "Properties of Proportional-Hazards Score Tests under Misspecified Regression Models," *Biometrics*, 40, 1037-1048.
- Lawless, J. F. (1987), "Regression Methods for Poisson Process Data," *The Journal of the American Statistical Association*, 82, 808-815.
- Lawless, J. F. and Nadeau, C. (1995), "Some Simple Robust Methods for the Analysis of Recurrent Events," *Technometrics*, 37, 158-168.
- Le Cam, L. (1956), "On the Asymptotic Theory of Estimation and Testing Hypotheses," *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability I*, University of California Press, 129-156.
- Lin, D. Y. and Wei, L. J. (1989), "The Robust Inference for the Cox Proportional Hazard Model," *The Journal of the American Statistical Association*, 84, 1074-1078.
- Montgomery, D. C. (1997), *Introduction to Statistical Quality Control, 3rd Ed.*, New York: Wiley.
- Nakamura, T. (1992), "Proportional Hazards Model with Covariates Subject to Measurement Error," *Biometrics*, 48, 829-838.

- Oakes, D. (1992), "Frailty Models for Multiple Event Times," *Survival Analysis: State of the Arts*, Ed. J. P. Klein and P. K. Goel, Netherlands: Kluwer Academic Publishers, pp. 371-9.
- Pepe, M. S., Self, S. G. and Prentice, R. L. (1989), "Further Results on Covariate Measurement Errors in Cohort Studies with Time to Response Data," *Statistics in Medicine*, 8, 1167-1178.
- Prentice, R. L. (1982), "Covariate Measurement Errors and Parameter Estimation in a Failure Time Regression Model," *Biometrika*, 69, 331-342.
- Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981), "On the Regression Analysis of Multivariate Failure Time Data," *Biometrika*, 68, 373-379.
- Raboud, J. M. (1991), *The Effects of Errors in Measurement in Survival Analysis*, Ph. D. Thesis, University of Toronto.
- Raboud, J., Reid, N., Coates, R. A. and Farewell, V. T. (1993), "Estimating Risks of Progressing to Aids when Covariates are Measured with Error," *Journal of the Royal Statistical Society A*, 156, 393-406.
- StataCorp (1997), *Stata Statistical Software, Release 5.0* Stata Corporation, College Station, Texas.
- Statistical Sciences Inc. (1995), *S-PLUS Statistical Software* Seattle, Washington.
- Taylor, J.R. (1997), *An Introduction to Error Analysis, 2nd Ed.*, Sausalito, California: University Science Books.
- Turnbull, B. W., Jiang, W. and Clark, L. C. (1997), "Regression Models for Recurrent Event Data: Parametric Random Effects Models with Measurement Error," *Statistics in Medicine*, 16, 853-864.
- Wei, L. J., Lin, D. Y. and Weissfeld, L. (1989), "Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions," *The Journal of the American Statistical Association*, 84, 1065-1073.
- White, H. (1994), *Estimation, Inference and Specification Analysis*, Cambridge University Press, Cambridge, England.

Table 1: Statistical analyses for several models of NPC trial SCC data

Model	Treatment			Baseline Se		
	estimate	(s.e.)	\mathcal{Z} -value	estimate	(s.e.)	\mathcal{Z} -value
1) Semi-parametric						
a) Naive	$\hat{g}=0.118$	(0.059)	2.014	—	—	—
b) Adjusted	$\hat{\gamma}=0.118$	(0.124)	0.954	—	—	—
c) Naive	$\hat{g}=0.117$	(0.059)	1.975	$\hat{b}=-0.690$	(0.146)	-4.737
d) Adjusted	$\hat{\gamma}=0.117$	(0.125)	0.933	$\hat{\beta}=-2.076$	(0.963)	-2.155
2) Constant Intensity						
a) Naive	$\hat{g}=0.124$	(0.059)	2.114	—	—	—
b) Adjusted	$\hat{\gamma}=0.124$	(0.124)	1.002	—	—	—
c) Naive	$\hat{g}=0.122$	(0.059)	2.058	$\hat{b}=-0.725$	(0.145)	-4.983
d) Adjusted	$\hat{\gamma}=0.122$	(0.125)	0.972	$\hat{\beta}=-2.181$	(0.963)	-2.264

Note: In (a), (b), only treatment is included in model. In (c),(d), Baseline Se is also included.

Captions for Figures

Figure 1: Estimate of baseline cumulative intensity function for SCC recurrence data for NPC trial. The dashed lines are the estimate and pointwise 95% confidence limits based on the naive analysis. The solid lines are the estimate and pointwise 95% confidence limits based on a robust analysis adjusted for random effects and covariate measurement error.

Figure 2: Normal probability plots and histograms of the estimated measurement errors $\{\hat{U}_i\}$ (in graphs labeled by U) and the estimated true log(Se) levels $\{\hat{X}_i\}$ (in graphs labeled by X) for 322 placebo group patients of the NPC trial with ten or more serial Se readings.