4-21-2010

Nonparametric and Semiparametric Analysis of Current Status Data Subject to Outcome Misclassification

Victor G. Sal y Rosas
University of Washington, gianceli@u.washington.edu

James P. Hughes
University of Washington, jphughes@u.washington.edu

1 Introduction

In many epidemiological studies, key objectives are (i) to estimate the distribution function (d.f.) of the time, T, to a particular event of interest, (ii) to test whether the d.f.s. of two groups are equal, and (iii) to measure the association of the failure time with a set of factors via a regression model. However, there are scenarios where T is not observed; instead one only observes whether or not T exceeds a random (or fixed by design) monitoring time Y. Data with this type of structure are known as current status data or type I interval-censored data.

Current status data arise in a variety of situations. For example, Diamond et al. (1986) studied the distribution of the age at weaning by evaluating whether or not an infant had been weaned at a single time after birth. Jewell and Shiboski (1990) described a study of HIV discordant couples where the HIV uninfected partner is tested at a single point in time to measure the incidence of HIV acquisition. Ferreira et al. (1996) estimated the distribution of duration of breastfeeding among Brazilian babies, using a single interview per mother, during routine pediatric consultations.

Groeneboom and Wellner (1992) characterized the nonparametric maximum likelihood estimator (NPMLE) of the d.f. of T, denoted by F, and some of its statistical properties for current status data. Van der Laan et al. (1997) studied a regularized NPMLE and proved that smooth functionals of this estimator are statistically efficient. Banerjee and Wellner (2005) proposed several methods to compute pointwise confidence intervals for the NPMLE of F such as: (a) using the asymptotic distribution of the NPMLE of F, (b) m out of n bootstrap and (c) inverting the likelihood ratio test of the hypothesis $H_0 : F(t_0) = \tau_0$ with $\tau_0 \in (0, 1)$. Similarly, several methods has been proposed for the two sample hypothesis testing problem: log rank type test (Sun, 1996; Sun and Kalbfleish, 1993, 1996; Sun, 1999), difference in survival means (Andersen and Ronn, 1995) and likelihood

ratio or score test based on specific types of alternative hypotheses (Kulikov, 2002). Huang (1996) studied the Cox proportional hazard model for current status data. Other regression models such as the proportional odds (Rossini and Tsiatis, 1996), acelerated failure time (Tian and Cai, 2006), linear (Shen, 2000), additive hazard (Lin et al., 1998) have also being studied.

An additional complication arises if the outcome of interest is measured imperfectly. For example, a test for disease may be insensitive and/or nonspecific; self report of weaning may be inaccurate due to social desirability bias; biopsies may miss a tumor; etc. In these cases, the methodology described above does not apply directly and one needs to account for the outcome misclassification in order to obtain a valid estimation of F. In the context of repeat testing, Balasubramanian and Lagakos (2001) estimated the risk of vertical transmission of HIV-1 assuming perfect specificity and time-dependent sensitivity. The same authors extended their ideas to the situation in which there could be different periods of exposure (see Balasubramanian and Lagakos, 2003). Richardson and Hughes (2000) implemented an EM algorithm to estimate the cumulative probability of disease in a discrete time context. Meier et al. (2003) extended their ideas using a Cox proportional model in a discrete-time context. Recently, McKeown and Jewell (2010) discussed adjusting for outcome misclassification under current status data. In particular, they described the NPMLE of F for the one sample problem, under misclassification, and extended their idea to a parametric regression setting.

A study conducted in Seattle, WA from 1998 to 2003 motivated our interest in this problem (Golden et al., 2005). The primary objective of the study was prevention of recurrent gonorrhea or chlamydial infection in patients 3 to 19 weeks after treatment and randomization to standard or expedited partner therapy. Patients in the expedited-treatment group were offered medication to give to their sex partners, of if they preferred, study staff members could contact their partners

and provided them with medication without a clinical examination. In this study, participants were observed only once during followup and their time of observation varied considerably. The test used to measure the outcome had low sensitivity (0.90) and good specificity (1.0).

McKeown and Jewell (2010) derive the NPLME for the distribution function of failure times but their regression modeling relied on parametric assumptions. In this article, we study more robust nonparametric and semiparametric methods. In section 2, we introduce needed notations and formulate the statistical problem. We then proceed to present inference results for the one sample problem, two sample hypothesis testing and semiparametric regression analysis. In section 3, we present simulation results and in section 4, an example using data from the aforementioned Partners Notification Study (Golden et al., 2005) is described. We conclude with a discussion and future directions of research.

2 Description of Data and Likelihood Function

2.1 Data structure

Assume that the failure time T is a random variable on $\mathbb{R}_+$ with d.f. F and Y is a random observation time on $\mathbb{R}_+$ with d.f. G. We observe only an indicator variable Δ that tells us whether the outcome has occurred ($\Delta = 1$) or not ($\Delta = 0$) at the observation time Y (i.e. $\Delta = 1_{[T \leq Y]}$). In addition, under outcome misclassification, we do not measure Δ directly; instead we observe an indicator variable $\tilde{\Delta}$ that is subject to misclassification. Denote the sensitivity and specificity of $\tilde{\Delta}$ by ϕ and ψ, respectively. More generally in a random sample the observation from the ith participant will be given by the vector $(Y_i, \tilde{\Delta}_i, \phi_i, \psi_i)$ where ϕ_i and ψ_i may vary among individuals. Finally, let $Y_{(i)}$ be the ith ordered value of $Y_1, \ldots, Y_n$ and $(\tilde{\Delta}_{(i)}, \phi_{(i)}, \psi_{(i)})$ are the indicator variable, sensitivity and specificity associated

with $Y_{(i)}$. There are two main assumptions that will hold throughout the paper (unless specified otherwise): (1) T is independent of Y and (2) ϕ and ψ are fixed and known with $\phi + \psi > 1$.

As noted in McKeown and Jewell (2010), Bayes's rule can be used to calculate the probability of observing a positive result ($\tilde{\Delta} = 1$) at the observation time Y as a function of ϕ, ψ and the true failure status at Y.

$$P(\tilde{\Delta} = 1 \mid Y) = \phi P(\Delta = 1 \mid Y) + (1-\psi)P(\Delta = 0 \mid Y) = 1 - \psi + (\psi + \phi - 1)F(Y) \tag{1}$$

More generally, sensitivity and specificity might vary at the subgroup or the individual level. In this context, we assume that all the values of sensitivity and specificity $\{\phi_i, \psi_i\}_{i=1}^n$ are known. Then, by condition (i) g does not have information about T, so the likelihood function for F, up to a constant, is given by

$$\begin{aligned} L_n(F) &= \prod_{i=1}^{n} \{1 - \psi_i + (\psi_i + \phi_i - 1)F(Y_i)]\}^{\tilde{\Delta}_i} \{\psi_i - (\psi_i + \phi_i - 1)F(Y_i)\}^{1-\tilde{\Delta}_i} \\ &= \prod_{i=1}^{n} \left\{1 - \psi_{(i)} + (\psi_{(i)} + \phi_{(i)} - 1)F(Y_{(i)})]\right\}^{\tilde{\Delta}_{(i)}} \times \\ &\quad \prod_{i=1}^{n} \left\{\psi_{(i)} - (\psi_{(i)} + \phi_{(i)} - 1)F(Y_{(i)})]\right\}^{1-\tilde{\Delta}_{(i)}} \end{aligned} \tag{2}$$

We wish to maximize the log likelihood function $l_n(F)$ ($l_n(.) = \log(L_n(.))$ over the space $\mathcal{F}$ defined as the space of right continuous increasing step functions, bounded by $[0, 1]$, with jumps at $Y_{(1)}, \ldots, Y_{(n)}$. Note that if the tests have perfect sensitivity and specificity, then $\tilde{\Delta} = \Delta$ and Groeneboom and Wellner (1992) provide an extensive study of this problem. We will denote the NPMLE of F as $\hat{F}_n$, and the naive estimator of the misclassified data assuming no misclassification as $\tilde{F}_n$.

Remark 1. If $\tilde{\Delta}_{(1)} = 0$ then the value of the NPMLE $\hat{F}_n$ at $Y_{(1)}$ can be set equal to zero without imposing additional constraints on the maximization problem. A

similar argument can be made if $\tilde{\Delta}_{(n)} = 1$ but in this case $\hat{F}_n(Y_{(n)})$ will be equal to one. Thus, without loss of generality we assume for the rest of the paper that $\tilde{\Delta}_{(1)} = 1$ and $\tilde{\Delta}_{(n)} = 0$.

2.2 Inferences

When all observations have the same sensitivity and specificity, the NPMLE of F_0 is given by the following proposition.

Proposition 1. (McKeown and Jewell, 2010) *The NPMLE of F at $Y_{(i)}$ is*

$$\hat{F}_n(Y_{(i)}) = \frac{\left\{\left[\tilde{F}_n(Y_{(i)}) \vee (1-\psi)\right] \wedge \phi\right\} + \psi - 1}{\phi + \psi - 1} \tag{3}$$

where $a \vee b = \max(a, b)$, $a \wedge b = \min(a, b)$, and $\tilde{F}_n$ is

$$\tilde{F}_n(Y_{(m)}) = \max_{i \leq m} \min_{k \geq m} \frac{\sum_{j=i}^{k} \tilde{\Delta}_{(j)}}{k - i + 1} \ , \ m \in \{1, \ldots, n\} \tag{4}$$

This proposition can be proven using similar arguments to those in the proof of Proposition 1.2 of Groeneboom and Wellner (1992) and a formal proof was recently presented by McKeown and Jewell (2010). Some statistical properties of this estimator are presented in the Web appendix.

In reality, sensitivity and specificity may vary across individuals or group of individuals. For instance, one may want to combine observations that were tested with different laboratory tests; due to budget considerations, a small proportion of the cohort may be tested with a more accurate test (possibly perfect sensitivity and specificity) and the remaining participants with a less accurate test. In these scenarios, it is not possible to express $\hat{F}_n$ explicitly as in (3) but it can still be characterized using the following proposition.

Proposition 2. *A point $\hat{\mathbf{x}} = (\hat{F}_n(Y_{(1)}), \ldots, \hat{F}_n(Y_{(n)}))$ is the NPMLE over the set*

$\{\mathbf{x} = (x_1, \ldots, x_n) \in (0,1)^n \ : x_1 \leq \cdots \leq x_n\}$ *if and only if* $\hat{\mathbf{x}}$ *is the left derivative of the convex minorant of the cumulative sum diagram of* $P_0 = (0,0)$ *and* $P_j = (G_j(\hat{\mathbf{x}}), V_j(\hat{\mathbf{x}}))$ *for* $j = 1, \ldots, n$ *where*

$$\begin{aligned} G_j(\mathbf{x}) &= \sum_{i=1}^{j} -\frac{\partial^2 l_n}{\partial x_i^2}(x_i) \\ &= \sum_{i=1}^{j} \left[\frac{\tilde{\Delta}_{(i)}(\phi_{(i)} + \psi_{(i)} - 1)^2}{[1 - \psi_{(i)} + (\psi_{(i)} + \phi_{(i)} - 1)x_i]^2} + \frac{(1 - \tilde{\Delta}_{(i)})(\phi_{(i)} + \psi_{(i)} - 1)^2}{[\psi_{(i)} - (\psi_{(i)} + \phi_{(i)} - 1)x_i]^2} \right] \end{aligned} \quad (5)$$

and

$$\begin{aligned} V_j(\mathbf{x}) &= \sum_{i=1}^{j} - \left[x_i - \left(\frac{\partial^2 l_n}{\partial x_i^2}(x_i) \right)^{-1} \frac{\partial l_n}{\partial x_i}(x_i) \right] \frac{\partial^2 l_n}{\partial x_i^2}(x_i) \\ &= \sum_{i=1}^{j} x_i \left[\frac{\tilde{\Delta}_{(i)}(\phi_{(i)} + \psi_{(i)} - 1)^2}{[1 - \psi_{(i)} + (\psi_{(i)} + \phi_{(i)} - 1)x_i]^2} + \frac{(1 - \tilde{\Delta}_{(i)})(\phi_i + \psi_{(i)} - 1)^2}{[\psi_{(i)} - (\psi_{(i)} + \phi_{(i)} - 1)x_i]^2} \right] \\ &\quad + \sum_{i=1}^{j} \left[\frac{\tilde{\Delta}_{(i)}(\psi_{(i)} + \phi_{(i)} - 1)}{1 - \psi_{(i)} + (\psi_{(i)} + \phi_{(i)} - 1)x_i} - \frac{(1 - \tilde{\Delta}_{(i)})(\psi_{(i)} + \phi_{(i)} - 1)}{\psi_{(i)} - (\psi_{(i)} + \phi_{(i)} - 1)x_i} \right] \end{aligned} \quad (6)$$

A proof of Proposition 2 is based on the discussion on Banerjee (2007, pp 9-10). Proposition 2 does not provide an explicit formula to compute $\hat{F}_n$; however, it suggests an iterative algorithm that is summarized as follows:

Algorithm 0.

(a) Set as an initial guess $\hat{\mathbf{x}}^{(0)}$ (e.g $\hat{x}_i^{(0)} = i/(n+1)$ for $i = 1, \ldots, n$).

(b) At stage k, compute the left derivative, $\hat{\mathbf{x}}^{(k)}$, of the convex minorant of the cumulative sum diagram formed by $P_0 = (0,0)$ and $P_j = \big(G_j(\hat{\mathbf{x}}^{(k-1)}), V_j(\hat{\mathbf{x}}^{(k-1)})\big)$, for $j = 1, \ldots, n$, according to

$$\hat{x}_m^{(k)} = \max_{i \leq m} \min_{l \geq m} \frac{V_l(\hat{\mathbf{x}}^{(k-1)}) - V_i(\hat{\mathbf{x}}^{(k-1)})}{G_l(\hat{\mathbf{x}}^{(k-1)}) - G_i(\hat{\mathbf{x}}^{(k-1)})} \quad , \quad m = 1, \ldots, n \qquad (7)$$

(c) Repeat step (b) until convergence e.g.

$$|\langle \nabla l_n(\hat{\mathbf{x}}^{(k)}), \hat{\mathbf{x}}^{(k)} \rangle| < \epsilon \quad , \quad \max \left\{ \sum_{i=k}^{s} \frac{\partial}{\partial x_i} l_n(\hat{\mathbf{x}}^{(k)}) \ : \ k = 1, 2, \ldots, s \right\} < \epsilon \tag{8}$$

then $\hat{F}_n = \hat{\mathbf{x}}^{(\infty)}$.

Jongbloed (1998) noted that this algorithm may not always converge and proposed a modified version, called the modified iterative convex minorant (MICM) algorithm, that guarantees global convergence. The MICM algorithm can be used in any maximization procedure where the log likelihood function is concave with respect to the parameters of interest and where those parameters have monotonicity constraints.

To make inferences, we compute a likelihood ratio based confidence interval by applying Theorem 2.2 in Banerjee (2007). The idea is to invert the complement of the rejection region for the hypothesis testing problem $H_0 : F(t_0) = \tau_0$ where $\tau_0 \in (0, 1)$ and $t_0 \in (0, \infty)$. A likelihood ratio statistic for testing H_0 and its asymptotic distribution is described below.

Proposition 3. *Suppose that F and G are continuously differentiable in a neighborhood of t_0 with $f(t_0) > 0$ and $g(t_0) > 0$, and assume that $\phi_i = \phi$ and $\psi_i = \psi$ for all observations. Denote the likelihood ratio statistic λ_n by*

$$\lambda_n(\tau_0) = \frac{L_n(\hat{F}_n)}{L_n(\hat{F}_n^0)} \tag{9}$$

where $\hat{F}_n^0$ is the NPMLE under H_0. Then the limiting distribution of the likelihood ratio statistic for testing H_0 is

$$2 \log \lambda_n(\tau_0) = 2[l_n(\hat{F}_n) - l_n(\hat{F}_n^0)] \rightarrow_d \mathcal{D} \tag{10}$$

$\mathcal{D}$ is a random variable that does not depend on F, G, ϕ, ψ or t_0 and a tabulation of the quantiles of this random variable is presented by Banerjee and Wellner (2001). Thus

$$C_{n,\alpha} = \{\tau \in (0,1) : 2\log\lambda_n(\tau) < d_\alpha\} \tag{11}$$

forms a $100(1-\alpha)\%$ confidence interval, where d_α is the $100(1-\alpha)$th percentile of $\mathcal{D}$. In practice, we present the following algorithm to compute $\hat{F}_n^0$.

Algorithm 1. (Same sensitivity and specificity for all observations)

(a) Find m such that $Y_{(m)} \le t_0 \le Y_{(m+1)}$

(b) For $\{Y_{(1)}, \ldots, Y_{(m)}\}$, compute the left derivative of the cumulative sum diagram (see eq 4) formed by $P_0 = (0,0)$ and $\left\{P_i = \left(i, \sum_{j=1}^{i} \tilde{\Delta}_{(i)}\right)\right\}_{i=1}^{m}$, denoted by $\eta = (\eta_1, \ldots, \eta_m)$. Then

$$\hat{F}_n^0(Y_{(i)}) = \left[\frac{\eta_i + \psi - 1}{\phi + \psi - 1} \vee 0\right] \wedge \tau_0 \tag{12}$$

for $i = 1, \ldots, m$.

(c) For $\{Y_{(m+1)}, \ldots, Y_{(n)}\}$, compute the left derivative of the cumulative sum diagram (see eq 4) formed by $P_0 = (0,0)$ and $\left\{P_i = \left(i, \sum_{j=1}^{i} \tilde{\Delta}_{(m+j)}\right)\right\}_{i=1}^{n-m}$, denoted by $\xi = (\xi_{m+1}, \ldots, \xi_n)$. Then

$$\hat{F}_n^0(Y_{(i)}) = \left[\frac{\xi_i + \psi - 1}{\phi + \psi - 1} \vee \tau_0\right] \wedge 1 \tag{13}$$

for $i = m+1, \ldots, n$.

Algorithm 2.(Varying sensitivity and specificity)

(a) Find m such that $Y_{(m)} \le t_0 \le Y_{(m+1)}$

(b) For $\{Y_{(1)}, \ldots, Y_{(m)}\}$, compute η that maximizes

$$\sum_{i=1}^{m} \left\{ \tilde{\Delta}_{(i)} \log\left[1 - \psi_{(i)} + (\psi_{(i)} + \phi_{(i)} - 1)x_i\right] + (1 - \tilde{\Delta}_{(i)}) \log\left[\psi_{(i)} - (\psi_{(i)} + \phi_{(i)} - 1)x_i\right] \right\}$$

over $\{\mathbf{x} \in (0,1)^m \ : \ x_1 \leq x_2 \leq \cdots \leq x_m\}$ using the MICM algorithm; then

$$\hat{F}_n^0(Y_{(i)}) = \eta_i \wedge \tau_0 \quad , \quad i = 1, \ldots, m \tag{14}$$

(c) For $\{Y_{(m+1)}, \ldots, Y_{(n)}\}$, compute ξ that maximizes

$$\sum_{i=m+1}^{n} \left\{ \tilde{\Delta}_{(i)} \log\left[1 - \psi_{(i)} + (\psi_{(i)} + \phi_{(i)} - 1)x_i\right] + (1 - \tilde{\Delta}_{(i)}) \log\left[\psi_{(i)} - (\psi_{(i)} + \phi_{(i)} - 1)x_i\right] \right\}$$

over $\{\mathbf{x} \in (0,1)^{n-m} \ : \ x_{m+1} \leq x_{m+2} \leq \cdots \leq x_n\}$ using the MICM algorithm; then

$$\hat{F}_n^0(Y_{(i)}) = \xi_i \vee \tau_0 \quad , \quad i = m+1, \ldots, n \tag{15}$$

Remark 2. Our approach to compute pointwise confidence intervals is computationally faster than the m out of n bootstrap idea proposed by McKeown and Jewell (2010). Moreover, as described above, it can be applied when sensitivity and specificity varies at the individual level and we study this with simulations.

2.3 Two Sample Hypothesis Testing

Consider a binary variable Z that denotes whether the person is in the "intervention"group ($Z = 1$) or the "control"group ($Z = 0$), and where the probability of being in the intervention group is denoted by p. Let F_0 and F_1 denote the d.f.s of the intervention and control groups respectively, and assume the observations times for both groups follow a d.f. G. Moreover, assume the sensitivity (ϕ) and

specificity (ψ) are the same for all observations. The following result suggests a natural statistic for testing $H_0 : F_0 = F_1$.

Proposition 4. *Suppose that*

(i) The support of F is a bounded interval $I = [0, M]$ with $G \ll F$, $F \ll G$, G has density g with respect to the Lebesgue measure, and h is a fixed measurable function.

(ii) F_0, g and h satisfy

$$I^{-1}(F, g, h) = \int_0^M \frac{[1 - \psi + (\psi + \phi - 1)F(y)][\psi - (\psi + \phi - 1)F(y)]}{(\psi + \phi - 1)^2 g(y)} h^2(y)dy < \infty \tag{16}$$

(iii) $(h/g) \circ F^{-1}$ is bounded and is a Lipschitz function on $[0, 1]$.

Then the functional $\nu(F) = \int_I (1 - F(t))h(t)dt$ satisfies

$$\sqrt{n}[\nu(\hat{F}_n) - \nu(F)] \to_d N(0, I^{-1}(F, g, h)) \tag{17}$$

Based on Proposition 4, we propose the following test statistic for H_0

$$U_n = \sqrt{\frac{n_1 n_0}{n}} \int_0^M [\hat{F}_{n_1}(t) - \hat{F}_{n_0}(t)]d\hat{G}_n \tag{18}$$

where $\hat{G}_n$ is the empirical d.f. of the observation times of the combined sample ($n = n_1 + n_0$), and $\hat{F}_{n1}, \hat{F}_{n0}$ are the NPMLE of F_1 and F_0 respectively. The limit distribution of the proposed statistic, under the null, is presented below.

Proposition 5. *Assuming the conditions of Lemma 3 hold and that $n_1/n \to a \in (0, 1)$ as $n \to \infty$. Then, under H_0*

$$U_n \to_d N(0, I^{-1}(F^{H_0}, g, g)) \tag{19}$$

where F^{H_0} is the common d.f under H_0.

Under the null hypothesis $\hat{U}_n/\sqrt{\hat{I}_n}$ can be approximated by a standard normal distribution, where

$$\hat{U}_n = \sqrt{\frac{n_1 n_0}{n^3}} \sum_{i=1}^{n} \left[\hat{F}_{n_1}(Y_{(i)}) - \hat{F}_{n_0}(Y_{(i)}) \right] \tag{20}$$

And

$$\hat{I}_n^{-1} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{[1-\psi+(\psi+\phi-1)\hat{F}_n^0(Y_{(i)})][\psi-(\psi+\phi-1)\hat{F}_n^0(Y_{(i)})]}{(\phi+\psi-1)^2} \right\} \tag{21}$$

Remark 3. Under the null hypothesis, as the sample size n increases the constraints on the naive estimator will become irrelevant (i.e. $P(\tilde{F}_n^0(t) < 1-\psi)$ and $P(\tilde{F}_n^0(t) > \phi)$ tend to zero as $n \to \infty$ if t is an interior point of the domain of F^{H_0}). Therefore, as n grows

$$\hat{F}_n^0 \approx \frac{\tilde{F}_n^0 + \psi - 1}{\phi + \psi - 1}$$

and

$$\frac{\tilde{U}_n}{\sqrt{\tilde{I}_n^{-1}}} \approx \frac{(\phi+\psi-1)\hat{U}_n}{(\phi+\psi-1)\sqrt{I_n^{-1}}} = \frac{\hat{U}_n}{\sqrt{\hat{I}_n^{-1}}}$$

where $\tilde{U}_n$ and $\tilde{I}_n^{-1}$ are naive estimators that assume no misclassification. Thus, when ϕ and ψ are constant across all individuals, misclassification can be ignored for testing $H_0 : F_0 = F_1$, as n goes to infinity. The behavior for finite sample sizes will be studied in section 7.

Remark 4. The asymptotic result presented above assumes constant sensitivity and specificity across all individuals; however equations, (20) and (21) suggest that the idea could be extended to individual level misclassification (which could not be ignored). We explore this potential with some simulations.

2.4 Semiparametric regression by the Cox proportional hazards model

The proportional hazard model is given by

$$\Lambda(Y|\mathbf{Z}) = \Lambda_0(Y)e^{(\mathbf{Z}'\theta)} \tag{22}$$

where Λ is the cumulative hazard at Y, and the covariate vector $\mathbf{Z} = (Z_1, \ldots, Z_r)$ is assumed to act additively on $\log(\Lambda(Y|\mathbf{Z}))$, Λ_0 is the baseline cumulative hazard independent of the covariates, and θ=$(\theta_1, \ldots, \theta_r)$ is the vector of log hazard ratios linking $\mathbf{Z}$. Since $F = 1 - e^{-\Lambda}$, we may combine (2) and (22), and the observed likelihood function, for an i.i.d sample of observations $(Y_1, \tilde{\Delta}_1, \mathbf{Z}_1), \ldots, (Y_n, \tilde{\Delta}_n, \mathbf{Z}_n)$, is proportional to

$$\begin{aligned} L_n(\theta, \Lambda) &= \prod_{i=1}^{n} \left\{ \phi_i - (\phi_i + \psi_i - 1)e^{-\Lambda(Y_i)e^{\mathbf{Z}_i'\theta}} \right\}^{\tilde{\Delta}_i} \\ &\quad \times \prod_{i=1}^{n} \left\{ 1 - \phi_i + (\phi_i + \psi_i - 1)e^{-\Lambda(Y_i)e^{\mathbf{Z}_i'\theta}} \right\}^{1-\tilde{\Delta}_i} \end{aligned} \tag{23}$$

where $\theta \in \Theta \subset R^r$ and $\Lambda \in \mathcal{G}$ where $\mathcal{G}$ is the set of nonnegative right-continuous increasing step functions (but bounded over the support of the observation time) with jump points at $Y_{(1)}, \ldots, Y_{(n)}$. In what follows, we denote the true underlying values of the parameters (θ, Λ) by (θ_0, Λ_0) and denote the maximum likelihood estimator by $(\hat{\theta}_n, \hat{\Lambda}_n)$.

Consider the problem of testing $H_0 : \theta = \theta_\mathbf{0}$ and define the likelihood ratio statistic

$$\lambda_n(\theta_0) = \frac{L_n(\hat{\theta}_n, \hat{\Lambda}_n)}{L_n(\theta_\mathbf{0}, \hat{\Lambda}_n^{\theta_\mathbf{0}})} \tag{24}$$

where $\hat{\Lambda}_n^{\theta_0}$ is the NPMLE of Λ_0 under H_0. The following proposition establishes the asymptotic distribution of $\hat{\theta}_n$ and $2\log(\lambda_n)$ when $\phi_i = \phi$ and $\psi_i = \psi$ for all $i = 1, \ldots, n$. The proof follows as an application of Theorem 3.1 in Banerjee et al. (2009).

Proposition 6. *Suppose that conditions (A.1)-(A.6) of Banerjee et al. (2009) hold and $\phi + \psi > 1$, then $\hat{\theta}_n$ is asymptotically linear in the efficient score function $\tilde{l}$, and has the representation*

$$\sqrt{n}(\hat{\theta}_n - \theta_{\mathbf{0}}) = \frac{1}{\sqrt{n}} I_0^{-1} \sum_{i=1}^{n} \tilde{l}(Y_i, \tilde{\Delta}_i, \mathbf{Z}_i) + o_P(1) \rightarrow_d N(0, I_0^{-1}) \tag{25}$$

Moreover

$$2\log(\lambda_n) = n(\hat{\theta}_n - \theta_{\mathbf{0}})^T I_0 (\hat{\theta}_n - \theta_{\mathbf{0}}) + o_P(1) \rightarrow_d \chi_r^2 \tag{26}$$

Proposition 6 can be used to develop likelihood ratio based confidence intervals for the regression coefficient. For example, consider the hypothesis $H_0 : \theta_p = \beta$, the likelihood ratio statistic is

$$\begin{aligned} 2\log\lambda_n(\beta) &= 2\log L_n(\hat{\theta}_n, \hat{\Lambda}_n) - 2\log L_n(\hat{\theta}_n^{-p}, \hat{\Lambda}_n^{-p}) \\ &= 2l_n(\hat{\theta}_n, \hat{\Lambda}_n) - 2l_n(\hat{\theta}_n^{-p}, \hat{\Lambda}_n^{-p}) \end{aligned} \tag{27}$$

where $(\hat{\theta}_n^{-p}, \hat{\Lambda}_n^{-p})$ is the NPMLE of the likelihood function (23) assuming $\theta_p = \beta$. Then, by Proposition 6, $2\log\lambda_n(\beta)$ has approximately a χ_1^2 distribution under H_0. As a consequence, $\{\beta : 2\log\lambda_n(\beta) \leq q_{1-\alpha}\}$ forms a $100(1-\alpha)\%$ confidence interval, where $q_{1-\alpha}$ is the $100(1-\alpha)$th percentile of the Chi-squared distribution with one degree of freedom.

One could also, in principal, consider using (25) to compute confidence intervals for the regression coefficients. However, that would involve estimation of additional nuisance parameters.

Remark 5. We do not specify the limit distribution of the estimated cumulative hazard function ($\hat{\Lambda}_n$). In most of the literature on semiparametric models with order restrictions on the nuisance parameter, the likelihood function is concave with respect to the nuisance paremeter. In those scenarios, finding the asymptotic behavior of $\hat{\Lambda}_n$ is possible using techniques from isotonic regression (Robertson et al., 1998) and convex optimization (Rockafellar, 1970). When current status data is subject to outcome misclassification, that concavity property does not always hold and depends on the value of the sensitivity (but not specificity). Therefore, the asymptotic behavior of the cumulative hazard function remains to be found. However, the main objective of this paper is to adjust for the baseline hazard to obtain an accurate estimation of the regression coefficient, therefore we postpone this problem for future research.

We now propose an algorithm to compute $(\hat{\theta}_n, \hat{\Lambda}_n)$. To avoid excessive notation we will consider the case of a single covariate ($r = 1$).

Algorithm 3. Estimation of the Regression Parameters

If the true disease status Δ is observed, then one estimate θ using the profile approach proposed by Huang (1996) or the joint maximization idea proposed by Pan (1999). If, instead, one observes $\tilde{\Delta}$ then we propose estimating θ using an EM algorithm (Dempster et al., 1977). Let $(\hat{\theta}_n^{(k)}, \hat{\Lambda}_n^{(k)})$ be the current estimates of (θ, Λ_0) at iteration k. Then the EM algorithm is:

(a) Expectation step:

$$\begin{aligned} Q\left[(\theta,\Lambda) \mid \hat{\theta}_n^{(k)}, \hat{\Lambda}_n^{(k)}\right] &= E_{\hat{\theta}_n^{(k)},\hat{\Lambda}_n^{(k)}}\left[l_n^C(\Delta_i, Y_i, Z_i) \mid (\tilde{\Delta}_i, Y_i, Z_i)\right] \\ &= E_{(k)}\left[l_n^C(\Delta_i, Y_i, Z_i) \mid (\tilde{\Delta}_i, Y_i, Z_i)\right] \\ &= l_n^C(E_{(k)}[\Delta_i \mid \tilde{\Delta}_i, Y_i, Z_i], Y_i, Z_i) \end{aligned} \tag{28}$$

where l_n^C is the log likelihood function for complete data (assuming that one

observes the true disease status Δ)

$$l_n^C(\theta, \Lambda) = \sum_{i=1}^{n} \left\{ \Delta_i \log \left[1 - e^{-\Lambda(Y_i)e^{Z_i\theta}} \right] - (1 - \Delta_i)\Lambda(Y_i)e^{Z_i\theta} \right\} \tag{29}$$

and

$$E_{(k)} \left[\Delta \mid \tilde{\Delta}, Y, Z \right] = \begin{cases} \dfrac{P^{(k)}[\Delta = 1 \mid Y, Z]\phi}{P^{(k)}[\Delta = 1 \mid Z, Y]\phi + P^{(k)}[\Delta = 0 \mid Z, Y](1 - \psi)} & , \quad \tilde{\Delta} = 1 \\ \dfrac{P^{(k)}[\Delta = 1 \mid Z, Y](1 - \phi)}{P^{(k)}[\Delta = 1 \mid Z, Y](1 - \phi) + P^{(k)}[\Delta = 0 \mid Z, Y)]\psi} & , \quad \tilde{\Delta} = 0 \end{cases} \tag{30}$$

where

$$P^{(k)}[\Delta = 1 \mid Y, Z] = 1 - \exp \left[-\hat{\Lambda}_n^{(k)}(Y) \exp \left(Z\hat{\theta}_n^{(k)} \right) \right] \tag{31}$$

(b) Maximization step: Update the parameters according to

$$\begin{aligned} (\hat{\theta}_n^{(k+1)}, \hat{\Lambda}_n^{(k+1)}) &= \arg \max_{\theta \in \Theta \ , \Lambda \in \mathcal{G}} Q \left[(\theta, \Lambda) \mid \hat{\theta}_n^{(k)}, \hat{\Lambda}_n^{(k)} \right] \\ &= \arg \max_{\theta \in \Theta \ , \Lambda \in \mathcal{G}} l_n^C(E_{(k)}[\Delta_i \mid \tilde{\Delta}_i, Y_i, Z_i], Y_i, Z_i) \end{aligned} \tag{32}$$

We alternative between the E and M steps until the following stopping criteria holds

$$\left| \frac{l_n(\hat{\theta}_n^{(k+1)}, \hat{\Lambda}_n^{(k+1)}) - l_n(\hat{\theta}_n^{(k)}, \hat{\Lambda}_n^{(k)})}{l_n(\hat{\theta}_n^{(k)}, \hat{\Lambda}_n^{(k)})} \right| \leq \epsilon \tag{33}$$

where ϵ is the tolerance level set in advance.

3 Simulation Studies

We conduct simulation studies to: (1) to assess the bias and misinterpretation of inference results when one ignores outcome misclassification and (2) to assess the

behavior of the proposed estimators for small sample sizes. These two objectives will be studied for different outcome prevalence, levels of misclassification and observation time distributions.

3.1 Simulations for the one sample problem

For the observation times, we consider the following distributions: continuous uniform, and exponential. For the distribution of failure time, we use a standard exponential distribution. We consider sample sizes of 500 and 1000 with 1000 simulations per scenario. We denote by p the expected proportion of observed failures and adjust the distribution of the observation times to achieve a fixed value of p ($p = P(T \leq Y)$). For $t_0 = G^{-1}(0.5)$, we compute the asymptotic percent bias ($\hat{b}_n$), defined as

$$\hat{b}_n = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{\hat{F}_n^{(r)}(t_0) - F_0(t_0)}{F_0(t_0)}\right] \tag{34}$$

and the nominal coverage of the 95% likelihood ratio-base confidence interval ($\hat{\gamma}_n$). Table 1 provides percent bias and coverage of selected estimators when the expected number of failures is 10%.

The coverage of the proposed confidence interval is very good and the average length of the confidence interval is shorter when 10% of the sample have been tested with a gold standard test (i.e. $\phi = \psi = 1$ for a random 10% of the observations). The bias of the naive estimator is affected the most by the specificity (as expected for low prevalence outcomes). Bias of the adjusted estimator is small and decreases as n increases while the unadjusted estimator remains biased regardless of n. Overall, the adjusted estimators have little bias and good coverage.

3.2 Simulations for two sample hypothesis testing

We assume that the failure time distribution in the control group is exponential with hazard rate equal to one ($\lambda_0 = 1$). The observation times follow a continuous uniform distribution for both groups, and p_0, p_1 are the expected proportion of observed failures for the control and intervention group, respectively. We compute the observed proportion of rejections under the null and proportional hazard alterantive hypothesis for situations of nondifferential and differential misclasification. Our simulations suggest (table 2) that for nondifferential misclassification, the adjusted test is more conservative than the unadjusted test statistic. However, as predicted by the asymptotic theory, as the sample sizes increases, this difference is diminished. Still, the unadjusted test behaves better in most of the studied scenarios when misclassification is not differential. The adjusted test is most conservative for low levels of specificity.

This behavior can be understood by noticing that, for low prevalence diseases and low specificity, equation 3 says that naive estimations lower than one minus specificity will be consider equal to zero by the NPMLE. That will mean that the variance estimator $\hat{I}_n$ (see equation 21) relays less on the data and more on the assume values of ϕ and ψ. That will induce the adjusted estimator to be more conservative in situations with low prevalence and low levels of specificity.

On the other hand, under differential misclassification, the adjusted test statistic preserves the correct type I error rate under the null hypothesis (table 3). In comparison, the unadjusted estimator is highly anticonservative due to the differential misclassification. As a consequence, when misclassification is not differential we recommend ignoring misclassification and computing a test statistics for the two sample test based on the unadjusted data. However, if misclassification is differential, then using the adjusted test is recommended.

3.3 Simulations for regression models

For our regression simulations, we assume that the baseline distribution function is a standard exponential ($\lambda_0 = 1$). We consider one or two binary covariates, each with probability of success equal to 0.5 and fix the sample size at 500 observations. We generate the observation times from a uniform distribution such that the expected number of left censored observations in the baseline group (for more than one covariate this group is defined by assigning all covariates equal to zero) was 10%. The number of replications was 1000 in all scenarios. Under those settings, we consider three models

(a) **Model 1:** $\Lambda(t) = t \exp(-0.695 Z_1)$ with $Z_1 \sim Bernoulli(0.5)$.

(b) **Model 2:** $\Lambda(t) = t \exp(0.405 Z_2)$ with $Z_2 \sim Bernoulli(0.5)$.

(c) **Model 3:** $\Lambda(t) = t \exp(-0.695 Z_1 + 0.405 Z_2)$ with Z_1, Z_2 independent and $Z_1, Z_2 \sim Bernoulli(0.5)$.

Table 4 shows that ignoring outcome misclassification induces attenuation of the association between the covariates and the failure time toward zero. Moreover, the higher the misclassification the stronger this attenuation is. For diseases with low prevalences (e.g HIV), we observed that the regression coefficients of the naive estimator are more affected by low levels of specificity than sensitivity.
Unexpectedly, the NPMLE of the regression coefficients (adjusting for misclassification) is biased upwards (based on the mean of the NPMLE across simulations). However, the median is closer to true value. In each of the study simulations, we observed that up to 5% of the generated datasets produced a regression coefficient that was far away from the true value. There was no evidence that these outlying estimates were driven by low number of events, the choice of the starting point or lack of convergence. Pan (1999) mentioned that the extended ICM, for Cox

regression in interval censored data (without misclassification), was slightly biased upwards (see Pan, 1999, page 116). The study of different algorithms and scenarios is a future area of research.

4 Application

The Partner Notification Study was conducted in King County Seattle, WA from September 1998 to March 2003 and enrolled heterosexual men and women who received a diagnosis of gonorrhea or genital chlamydia (Golden et al., 2005). Researchers contacted clinicians who diagnosed the infections to seek permission to contact their participants and to minimize the likelihood of reinfection after treatment but before randomization, patients who could not be contacted within 14 days after treatment were not eligible for the study. Each participant was randomized to expedited partner treatment (intervention) or standard partner referral (control). The primary outcome was persistent or recurrent gonorrhea and/or chlamydial infection (we will consider a composite outcome only) in the original participant at 90 days after enrollment although actual follow up times varied considerably due to difficulty contacting participants and scheduling followup visits. Of the 1864 participants, 931 were randomized to the intervention and 933 to the control group. A high proportion of participant were women (80%), the median age was 21 years and a large number of participants who were treated for chlamydia were enrolled in the study (see table 5); however, all these characteristics were balanced in each arm. The sensitivity and specificity of the test used to diagnose gonorrhea and/or chlamydia were approximately 0.9 and 1 respectively. In order to avoid missing an infection that could happened between enrollment and the observation time, participants were asked whether they repeated their treatment using medication intended for a partner; only one person acknowledge doing so. The

observation times was similar in both groups with median 87 days (IQR: 77-103) for the control group and 87 days (IQR: 76-104) for the intervention group.
We compute the NPMLE separately for the control and intervention group. This is presented in Figure 1. The application of the proposed two sample hypothesis test gave a p-value of 0.024. In a univariate analysis, participants in the intervention group were 26% less likely of reinfection than participants in the control group (HR=0.744 [95%CI: 0.580-0.956], p-value = 0.031). In a multivariate analysis, adjusting for gender and the interaction of gender and intervention, there was no evidence that the effect of the intervention was different for men and women (p-value of the iteraction = 0.331). In conclusion, participants in the control group have significant higher risk of recurrence of gonorrhea and/or chlamydial infection.

5 Discussion and Future Research

Most laboratory tests that are used to diagnose diseases have sensitivity and specificity less than one (e.g culture for gonorrhea, sputum-smear test for TB). On average, if the observed disease prevalence during the study followup is low (high) and specificity (sensitivity) is less than 1 then the use of the standard methodology that does not account for outcome misclassification, results in overestimation (underestimation) of the cumulative probability of failure.
We develop methodology for hypothesis testing and semiparametric regression that account for outcome misclassification in current status data by extending existing models that assume no misclassification. For the regression problem, we choose the Cox proportional hazard model because of its popularity. However, the same ideas can be used to extend other regression models (see Banerjee et al., 2009). In some situations (e.g predicting survival), one would like to make inference on

the cumulative hazard function when considering a regression model. Finding the correct limit distribution of the cumulative hazard function remains an open problem.

There is an important limitation in the analysis of our application example. A number of participants reported symptoms at their follow up visit (n=408, 22.3%) but this percentage did not differ meaningfully between the groups (21.1 % in the placebo group vs. 23.5% in the treatment group). This feature of the data may suggest a violation of the assumption of independence of the failure and observation times. However, we also analyzed the data after deleting the symptomatic cases and the results were qualitatively similar (data not shown). Estimation and inference in this setting (adding information on symptoms) is under investigation. In the hypothesis testing problem, we only consider situations when one can assume that the observation times are not different between groups. It is reasonable to assume that with some work our methods can also be extended to that situation where the observation times differ between groups.

The methodology presented in this paper assumes perfect knowledge of the values of ϕ and ϕ, and, in most cases, that they are the same for all observations. However, our simulations suggest that many of our results can be extended to situations were sensitivity and specificity vary across individuals or subgroups. It is of interest to further study from the theoretical perspective whether the two sample hypothesis testing and regression ideas can in all cases be extended to situations where the sensitivity and specificity vary. Another important potential extension is to incorporate uncertainty about ϕ and ψ. Finally, more efficient algorithms to speed up the computational time, especially for the regression analysis, need to be studied.

6 Appendix

For all the proofs considered in this web appendix, we will assume that all observations were subject to the same test with sensitivity (ϕ) and specificity (ψ) unless specified otherwise.

6.1 One sample problem

Proof of Proposition 2. The proof of this proposition can be adapted from page 8 and 9 in Banerjee (2007) or following the arguments to prove Theorem 4.3. in Wellner and Zhan (1997).

Definition 1 (Schick and Yu, 2000). *We say that t_0 is a regular point if $G(t_0 + \epsilon) - G(t_0) > 0$ and $G(t_0) - G(t_0 - \epsilon) > 0$ for every $\epsilon > 0$.*

Proposition A.1. *Let G be the distribution of the observation then*

I. The NPMLE $\hat{F}_n$ satisfies

$$\int \mid \hat{F}_n - F \mid dG \rightarrow_{a.s} 0 \tag{35}$$

II. Define

$$\Omega_G = \left\{ \omega \ : \ \int \mid \hat{F}_n - F \mid dG \rightarrow 0 \right\}$$

If t_0 is a regular continuity point of F_0 then for each $\omega \in \Omega_G$,

$$\hat{F}_n(t_0, \omega) \rightarrow F(t_0) \tag{36}$$

III. Suppose that F is continuous and that for all $a < b$, $0 < F(a) < F(b) < 1$ implies $0 < G(a) < G(b)$ then

$$\sup_{t \in [0,\infty)} \mid \hat{F}_n(t) - F(t) \mid \rightarrow_{a.s} 0 \tag{37}$$

Proof of Proposition A.1.

Part I. Define $\mu = \# \times G$ where $\#$ is the counting measure on $\{0, 1\}$. Then, the density of $(\tilde{\Delta}, Y)$ with respect to μ is given by

$$p_F(\tilde{\Delta}, Y) = [1 - \psi + (\phi + \psi - 1)F(Y)]^{\tilde{\Delta}}[\psi - (\phi + \psi - 1)F(Y)]^{1-\tilde{\Delta}}$$

One can see that the envelope function q of $\mathcal{F} = \{p_F \ : \ F \ d.f \ on \ \mathbb{R}^+\}$ is bounded by ϕ.

$$q = \sup_{p_F \in \mathcal{F}} p_F \leq \phi \leq 1$$

and one can apply Lemma 3.8 in Van de Geer (2000) to show that the δ-entropy with bracketing of $\mathcal{F}$ is bounded: $H_B(\delta, \mathcal{F}, \mu) \leq A\delta^{-1}$ for some constant A and all $\delta > 0$ (see Van de Geer, 2000, for definition of entropy). Therefore, by Lemma 4.4 in Van de Geer (2000) one has

$$h^2(\hat{p}_n, p_F) \rightarrow_{a.s} 0$$

where h^2 is the Hellinger metric. Also, since the total variation distance is dominated by the Hellinger metric, one has that

$$d_{TV}(p_F, p_{\hat{F}_n}) = 2 \mid \phi + \psi - 1 \mid \int |F - \hat{F}_n| dG \rightarrow_{a.s} 0 \tag{38}$$

Part II and III. The proofs of **II** and **III** are consequences of I and Propositions 1 and 4 respectively from Schick and Yu (2000).

The following Proposition, that includes Proposition 3 in the paper, is a consequence of applying Theorems 2.1 and 2.1 in Banerjee (2007).

Proposition A.2 *For any t such that F and G are continuous and differentiable, with densities f and g respectively, in a neighborhood of t with $g(t) > 0$, $f(t) > 0$,*

and $F(t) \in (0,1)$ *then*

$$n^{1/3}[\hat{F}_n(t) - F(t)] \to_d C\mathbb{Z} \tag{39}$$

where $\mathbb{Z} = \arg\min\{\mathbb{W}(t) + t^2\}$ *and* $\mathbb{W}$ *is a two-sided Brownian motion starting from* 0 *and*

$$C = \left\{ \frac{4[1-\psi+(\psi+\phi-1)F(t)][\psi-(\psi+\phi-1)F(t)]f(t)}{(\phi+\psi-1)^2 g(t)} \right\}^{1/3}$$

Moreover

$$2\log\lambda_n \to \mathcal{D} \qquad \textit{when } H_0 \textit{ is true}$$

Proof of Proposition A.2. The functional of interest is $\theta = F(Y)$ with density

$$p_\theta(\tilde{\Delta}, Y) = [1-\psi+(\psi+\phi-1)\theta]^{\tilde{\Delta}}[\psi-(\psi+\phi-1)\theta]^{(1-\tilde{\Delta})} g(Y) \tag{40}$$

and

$$I(\theta) = -E_\theta\left[\frac{\partial^2 \log(p_\theta)}{\partial\theta^2}\right] = \frac{(\psi+\phi-1)^2}{[1-\psi+(\psi+\phi-1)\theta][\psi-(\psi+\phi-1)\theta]} \tag{41}$$

Then the result follows from Theorem 2.1 in Banerjee (2007) with

$$a = \left\{ \frac{(\psi+\phi-1)^2 g(t)}{[1-\psi+(\psi+\phi-1)F(t)][\psi-(\psi+\phi-1)F(t)]} \right\}^{-1/2} \tag{42}$$

and $b = \frac{1}{2}f(t)$.

6.2 Hypothesis testing

Proof of Proposition 4. Notice that

$$\begin{aligned}\nu(F) &= \int_I [1-F(y)]h(y)dy = E_Y\left\{\frac{[1-F(Y)]h(Y)}{g(Y)}\right\} = E_Y\left\{\frac{\left[1-\frac{E[\tilde{\Delta}|Y]+\psi-1}{\phi+\psi-1}\right]h(Y)}{g(Y)}\right\} \\ &= E\left[\frac{\phi-\tilde{\Delta}}{(\phi+\psi-1)}\frac{h(Y)}{g(Y)}\right] = \int\left[\frac{\phi-\tilde{\Delta}}{(\phi+\psi-1)}\frac{h(y)}{g(y)}\right]dP(t,y) \qquad (43)\end{aligned}$$

Then

$$\begin{aligned}\sqrt{n}[\nu(\hat{F}_n)-\nu(F)] &= \sqrt{n}\int_I\left\{1-\hat{F}_n(y)-\frac{\phi-\tilde{\Delta}}{(\phi+\psi-1)}\right\}\frac{h(y)}{g(y)}dP(t,y) \\ &= \sqrt{n}\int_I\left\{\frac{\tilde{\Delta}-[1-\psi+(\psi+\phi-1)\hat{F}_n(y)]}{(\phi+\psi-1)}\right\}\frac{h(y)}{g(y)}dP(t,y)\end{aligned}$$

Moreover

$$\begin{aligned}\sqrt{n}[\nu(\hat{F}_n)-\nu(F)] &= \sqrt{n}\int_I\left\{\frac{\tilde{\Delta}-[1-\psi+(\psi+\phi-1)\hat{F}_n(y)]}{(\phi+\psi-1)}\right\}\frac{h(F^{-1}(\hat{F}_n(y)))}{g(F^{-1}(\hat{F}_n(y)))}dP(t,y)+ \\ &\quad \sqrt{n}\int_I\left\{\frac{\tilde{\Delta}-[1-\psi+(\psi+\phi-1)\hat{F}_n(y)]}{(\phi+\psi-1)}\right\}\frac{h(y)}{g(y)}dP(t,y) \\ &\quad -\sqrt{n}\int_I\left\{\frac{\tilde{\Delta}-[1-\psi+(\psi+\phi-1)\hat{F}_n(y)]}{(\phi+\psi-1)}\right\}\frac{h(F^{-1}(\hat{F}_n(y)))}{g(F^{-1}(\hat{F}_n(y)))}dP(t,y) \\ &= \sqrt{n}\int_I\left\{\frac{\tilde{\Delta}-[1-\psi+(\psi+\phi-1)\hat{F}_n(y)]}{(\phi+\psi-1)}\right\}\frac{h(F^{-1}(\hat{F}_n(y)))}{g(F^{-1}(\hat{F}_n(y)))}dP(t,y)+ \\ &\quad \sqrt{n}\int_I[F(y)-\hat{F}_n(y)]\left\{\frac{h(y)}{g(y)}-\frac{h(F^{-1}(\hat{F}_n(y)))}{g(F^{-1}(\hat{F}_n(y)))}\right\}dG(y) \\ &= -I_1+I_2 \qquad (44)\end{aligned}$$

Let $F^{-1}(u)=\inf\{t \ : \ F(t)\geq u\}$. Then, one can define $\eta=(h/g)\circ F^{-1}$. Then, following the ideas describe in page 43 (in particular equations 1.19 and 1.20) in

Groeneboom and Wellner (1992), or page 160 (see equations 3 and 4) in Huang and Wellner (1995), one has

$$\int_I \left\{ \frac{\tilde{\Delta} - [1 - \psi + (\psi + \phi - 1)\hat{F}_n(y)]}{(\phi + \psi - 1)} \right\} \eta(\hat{F}_n(y)) dP_n(t, y) = 0 \tag{45}$$

Thus

$$\begin{aligned}
-I_1 &= \sqrt{n} \int_I \left\{ \frac{\tilde{\Delta} - [1 - \psi + (\psi + \phi - 1)\hat{F}_n(y)]}{(\phi + \psi - 1)} \right\} \eta(\hat{F}_n(y)) d(P_n - P)(t, y) \\
&= \sqrt{n} \int_I \left\{ \frac{\tilde{\Delta} - [1 - \psi + (\psi + \phi - 1)F(y)]}{(\phi + \psi - 1)} \right\} \eta(\hat{F}(y)) d(P_n - P)(t, y) \\
&- \sqrt{n} \int_I [\hat{F}_n(y) - F(y)] \eta(\hat{F}(y)) d(P_n - P)(t, y) \\
&= \sqrt{n} \int_I \left\{ \frac{\tilde{\Delta} - [1 - \psi + (\psi + \phi - 1)F(y)]}{(\phi + \psi - 1)} \right\} \eta(F(y)) d(P_n - P)(t, y) + \\
&\sqrt{n} \int_I \left\{ \frac{\tilde{\Delta} - [1 - \psi + (\psi + \phi - 1)F(y)]}{(\phi + \psi - 1)} \right\} \left\{ \eta(\hat{F}_n(y)) - \eta(F(y)) \right\} d(P_n - P)(t, y) \\
&- \sqrt{n} \int_I [\hat{F}_n(y) - F(y)] \eta(\hat{F}(y)) d(P_n - P)(t, y) \\
&= I_{11} + I_{12} + I_{13}
\end{aligned} \tag{46}$$

I_{12} and I_{13} are $o_p(1)$ by arguments presented in Huang and Wellner (1995). Moreover

$$\begin{aligned}
I_{11} &= \sqrt{n} \int_I \left\{ \frac{\tilde{\Delta} - [1 - \psi + (\psi + \phi - 1)F(y)]}{(\phi + \psi - 1)} \right\} \frac{h(y)}{g(y)} d(P_n - P)(t, y) \\
&= -\sqrt{n}(P_n - P)(\tilde{l})
\end{aligned} \tag{47}$$

where

$$\tilde{l}(y, \tilde{\Delta}) = -\frac{1}{\phi + \psi - 1} \left\{ \tilde{\Delta} - [1 - \psi + (\psi + \phi - 1)F(y)] \right\} \frac{h(y)}{g(y)} 1_{[y>0]} \tag{48}$$

is the efficiency influence function for the population mean. Therefore

$$\sqrt{n}[\nu(\hat{F}_n) - \nu(F)] = \sqrt{n}(P_n - P)(\tilde{l}) + o_P(1) \rightarrow_d N(0, I^{-1}(F, g, h)) \tag{49}$$

and

$$\begin{aligned} I^{-1}(F,g,h) &= E[\tilde{l}^2(Y,\tilde{\Delta})] = E\left[\frac{\left\{\tilde{\Delta} - [1-\psi+(\psi+\phi-1)F(Y)]\right\}^2 h^2(Y)}{(\phi+\psi-1)^2 g^2(Y)}\right] \\ &= \int_0^M \left\{\frac{[1-\psi+(\psi+\phi-1)F(y)][\psi-(\psi+\phi-1)F(y)]h^2(y)}{(\phi+\psi-1)^2 g(y)}\right\} dy \end{aligned}$$

Proof of Proposition 5. Notice that

$$\begin{aligned} U_n &= \sqrt{\frac{n_1 n_0}{n}} \int_0^M [\hat{F}_{n1}(y) - \hat{F}_{n0}(y)] d\hat{G}_n(y) \\ &= \sqrt{\frac{n_1 n_0}{n}} \int_0^M [\hat{F}_{n1}(y) - F^{H_0}(y)] d\hat{G}_n(y) + \sqrt{\frac{n_1 n_0}{n}} \int_0^M [\hat{F}_{n1}(y) - F^{H_0}(y)] d\hat{G}_n(y) \\ &\rightarrow_d \sqrt{a}\, N(0, I^{-1}(F^{H_0}, g, g)) + \sqrt{1-a}\, N(0, I^{-1}(F^{H_0}, g, g)) =_d N(0, I^{-1}(F^{H_0}, g, g)) \end{aligned}$$

6.3 Regression Model

The following are the main conditions needed to prove Proposition 6. These are the same conditions specified by Banerjee et al. (2009). We modify our notation and consider Y the random variable and y an observation of that r.v.

Main assumptions for Proposition 6.

(A1) θ_0 is an interior point of $\Theta \subset R^k$, where Θ is a bounded subset.

(A2) The covariate Z has bounded support which means that exist z_0 such that $P(|Z| \leq z_0) = 1$. Also $E\{Var(X|Y)\}$ is positive definite with probability one.

(A3) $\Lambda_0(0) = 0$ and let $\kappa_{\Lambda_0} = \inf\{y \ : \ \Lambda_0(y) = \infty\}$. Then, the support of Y is an

interval $S[Y] = [\kappa_Y, \zeta_Y]$ with $0 < \kappa_Y \leq \zeta_Y < \kappa_{\Lambda_0}$.

(A4) $0 < \Lambda_0(\kappa-) < \Lambda_0(\zeta) < M$ where M is some large constant. Also Λ_0 is continuously differentiable on $S[Y]$ with derivative λ_0 bounded away from 0 and ∞.

(A5) The marginal density of Y, denoted by g_Y is continuous and positive on $S[Y]$.

(A6) The function a_* defined below has a version which is differentiable componentwise with each component having a bounded derivative on $S[Y]$.

(A7) Sensitivity and specificity satisfy $\phi + \psi > 1$.

Proof of Proposition 6. We will present the calculation of the efficient score function for the regression coefficient and as a consequence the information matrix for a real value covariate (The multivariate case follows easily from there). The rest of the proof follows from Banerjee et al. (2009).
The score function for θ is

$$\dot{l}_\theta = (\phi + \psi - 1) z \Lambda(y|z) S(y|z) Q(y, \tilde{\Delta}, z) \tag{50}$$

On the other hand, let $\mathcal{F}_0 = \{F_\eta \ : \ |\eta| < 1\}$ is a regular parametric subfamily of $\mathcal{F} = \{F \ : \ F \ll \mu\}$ where μ is the Lebesgue measure and

$$a(y) = \frac{\partial}{\partial \eta} \log f_\eta(y) \mid_{\eta=0}$$

where f_0 is the density of F and f_η is a one dimensional smooth curve through f. By definition $a \in L_2^0(F)$ where

$$L_0^2(F) = \{a \ : \ \int a dF = 0 \ , \ \int a^2 dF < \infty\}$$

Note that

$$\begin{aligned}\frac{\partial}{\partial\eta}S_\eta(y)\Big|_{\eta=0} &= \frac{\partial}{\partial\eta}\int_y^\infty dF_\eta(y)\Big|_{\eta=0} = \int_y^\infty \frac{\partial}{\partial\eta}dF_\eta(y)\Big|_{\eta=0} \\ &= \int_y^\infty \frac{\partial}{\partial\eta}\log f_\eta(y)\Big|_{\eta=0} dF = \int_y^\infty a dF \end{aligned} \tag{51}$$

then the score operator for f is

$$\dot{l}_f(a) = \frac{\partial}{\partial\eta}l(\theta, S_\eta)\Big|_{\eta=0} = \frac{\tilde{\Delta}(\phi+\psi-1)\ \frac{\partial}{\partial\eta}F_\eta(y|z)\Big|_{\eta=0}}{1-\psi+(\psi+\phi-1)F(Y|Z)} - \frac{(1-\tilde{\Delta})(\phi+\psi-1)\ \frac{\partial}{\partial\eta}F_\eta(y|z)\Big|_{\eta=0}}{\psi-(\psi+\phi-1)F(y|z)}$$

where

$$\begin{aligned}\frac{\partial}{\partial\eta}F_\eta(y|z)\Big|_{\eta=0} &= -\frac{\partial}{\partial\eta}S_\eta(y|z)\Big|_{\eta=0} = -\frac{\partial}{\partial\eta}S_\eta(y)^{\exp(\theta z)}\Big|_{\eta=0} \\ &= -\frac{\exp(\theta z)S(y|z)}{S(y)}\int_y^\infty a dF \end{aligned} \tag{52}$$

then the score operator for f is

$$\begin{aligned}\dot{l}_f(a) &= -(\phi+\psi-1)\frac{\exp(\theta z)S(y|z)}{S(y)}\int_y^\infty a dF \times \\ &\quad \left[\frac{\tilde{\Delta}}{1-\psi+(\psi+\phi-1)F(Y|Z)} - \frac{(1-\tilde{\Delta})}{\psi-(\psi+\phi-1)F(y|z)}\right] \\ &= -(\phi+\psi-1)\frac{\exp(\theta z)S(y|z)Q(y,\tilde{\Delta},z)}{S(y)}\int_y^\infty a dF \end{aligned} \tag{53}$$

In order to determine the efficient score l_θ^* for θ, one needs to find a_* such that

$$E\left\{[\dot{l}_\theta - \dot{l}_f a_*]\dot{l}_f a\right\} = 0 \quad , \quad \forall\ a\ \in L_2^0(F)$$

then

$$
\begin{aligned}
-\frac{E\left\{[\dot{l}_\theta - \dot{l}_f a_*]\dot{l}_f a\right\}}{(\phi+\psi-1)^2} &= E\left\{\exp(2\theta Z)S(Y|Z)^2 Q(Y,\tilde{\Delta},Z)^2\left[Z\Lambda + \frac{\int_Y^\infty a_* dF}{S(Y)}\right]\frac{\int_Y^\infty a dF}{S(Y)}\right\} \\
&= E\left\{\frac{\int_Y^\infty a dF}{S(Y)} E\left[\exp(2\theta Z)S(Y|Z)^2 Q(Y,\tilde{\Delta},Z)^2\left[Z\Lambda + \frac{\int_Y^\infty a_* dF}{S(Y)}\right] \mid Y\right]\right\} \\
&= E\left\{\frac{\int_Y^\infty a dF}{S(Y)} E\left[\exp(2\theta Z)O(Y|Z)\left[Z\Lambda + \frac{\int_Y^\infty a_* dF}{S(Y)}\right] \mid Y\right]\right\}
\end{aligned}
$$

then

$$
\Lambda(Y)E\left[\exp(2\theta Z)O(Y|Z)Z \mid Y\right] = -\frac{\int_Y^\infty a_* dF}{S(Y)} E\left\{\exp(2\theta Z)O(Y|Z) \mid Y\right\}
$$

and that implies that

$$
\int_Y^\infty a_* dF = -\frac{\Lambda(Y)S(Y)E\left[\exp(2\theta Z)O(Y|Z)Z \mid Y\right]}{E\left[\exp(2\theta Z)O(Y|Z) \mid Y\right]}
$$

Thus the efficient score function for θ is

$$
\begin{aligned}
\dot{l}^*_\theta &= \dot{l}_\theta - \dot{l}_f a_* \\
&= (\phi+\psi-1)\Lambda(Y)\exp(\theta Z)S(Y \mid Z)Q(Y,\tilde{\Delta},Z)\left[Z - \frac{E\left[Z\exp(2\theta Z)O(Y|Z) \mid Y\right]}{E\left[\exp(2\theta Z)O(Y|Z) \mid Y\right]}\right]
\end{aligned}
$$

and the information matrix is

$$
I(\theta) = E\left\{l^*_\theta\right\}^2 = (\phi+\psi-1)^2 E\left\{\Lambda^2(Y|Z)O(Y|Z)\left[Z - \frac{E\left[Z\exp(2\theta Z)O(Y|Z) \mid Y\right]}{E\left[\exp(2\theta Z)O(Y|Z) \mid Y\right]}\right]^2\right\}
$$

References

Andersen, P. K. and Ronn, B. B. (1995). A nonparametric test for comparing two samples where all observations are either left- or right-censored. *Biometrics* **51,**

323–329.

Balasubramanian, R. and Lagakos, S. W. (2003). Estimation of a failure time distribution based on imperfect diagnostic test. *Biometrika* **90,** 171–182.

Balasubramanian, S. and Lagakos, S. W. (2001). Estimation of the timing of perinatal transmission of HIV. *Biometrics* **57,** 1048–1058.

Banerjee, M. (2007). Likelihood based inference for monotone response models. *The Annals of Statistics* **35,** 931–956.

Banerjee, M., D., M., and Mishra, S. (2009). Semiparametric binary regression models under shape constraints with application to Indian schooling data. *Journal of Econometrics* **249,** 101–117.

Banerjee, M. and Wellner, J. (2001). Likelihood ratio tests for monotone functions. *The Annals of Statistics* **29,** 1699–1731.

Banerjee, M. and Wellner, J. (2005). Confidence intervals for current status data. *Board of the Foundation of the Scandinavian Journal of Statistics* **32,** 405–424.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B(Methodological)* **39,** 1–38.

Diamond, I., McDonald, J., and Shah, I. (1986). Proportional hazard models for current status data: Application to the study of differentials in age at weaning in pakistan. *Demography* **23,** 607–620.

Ferreira, M., Cardoso, M., Santos, A., Ferreira, C., and Szarfarc, S. (1996). Rapid epidemiologic assessment of breastfeeding practices: Probit analysis of current status data. *Journal of Tropical Pediatrics* **42,** 50–53.

Golden, M., Whittington, W., Handsfield, H., Hughes, J., Stamm, W., Hogben, M., Clark, A., Malinski, C., Helmers, J., Thomas, K., and Holmes, K. (2005). Effect of expedited treatment of sex partners on recurrent or persistent gonorrhea or chlamydial infection. *New England Journal of Medicine* **352,** 676–685.

Groeneboom, P. and Wellner, J. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhauser Verlag.

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

Huang, J. and Wellner, J. (1995). Asymptotic normality of the NPMLE of linear functionals for interval censored data, case 1. *Statistica Neerlandica* **49,** 153 – 163.

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

Jongbloed, G. (1998). The iterative convex minorant algorithm for nonparametric estimation. *Journal of Computational and Graphical Statistics* **7,** 310 – 321.

Kulikov, V. N. (2002). *Direct and Indirect Use of Maximum Likelihood*. PhD thesis, Delft University.

Lin, D., Oakes, D., and Ying, Z. (1998). Additive hazard regression with current status data. *Biometrika Trust* **85,** 289 – 298.

McKeown, K. and Jewell, N. P. (2010). Misclassification of current status data. *Lifetime Data Analysis* **16,** 215–230.

Meier, A., Richardson, B., and Hughes, J. P. (2003). Discrete proportional hazard models for mismeasured outcomes. *Biometrics* **59,** 947 – 954.

Pan, W. (1999). Extending the iterative convex minorant algorithm to the cox model for interval-censored Data. *Journal of Computational and Graphical Statistics* **8,** 109–120.

Richardson, B. A. and Hughes, J. P. (2000). Product limit estimation for infections diseases data when the diagnostic test for the outcome is measured with uncertainty. *Biostatistics* **1,** 341–354.

Robertson, T., Wright, F. T., and Dykstra, R. (1998). *Order restricted statistical inference*. Chichester ; New York : Wiley.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton Univ. Press.

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

Schick, A. and Yu, Q. (2000). Consistency of the GMLE with mixed case interval-censored data. *Scandinavian Journal of Statistics* **27,** 45 – 55.

Shen, X. (2000). Linear regressioin with current status data. *Journal of the American Statistical Association* **95,** 842 – 852.

Sun, J. (1996). A nonparametric test for interval-censored failure time data with application to AIDS studies. *Statistics in Medicine* **15,** 1387 – 1395.

Sun, J. (1999). A nonparametric test for current status data with unequal censoring. *Journal of the Royal Statistical Society: Series B* **61,** 243 – 250.

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

Sun, J. and Kalbfleish, J. D. (1996). Nonparametric test of tumor prevalence data. *Biometrics* **52,** 726 – 731.

Tian, L. and Cai, T. (2006). On the accelerated failure time model for current status and interval censored data. *Biometrika* **93,** 329 – 342.

Van de Geer, S. (2000). *Empirical Process in M-Estimation.* Cambridge Series in Statistical and Probabilistic Mathematics.

Van der Laan, M. J., Bickel, P. J., and Jewell, N. P. (1997). Singly and doubly censored current status data: Estimation, asymptotics and regression. *Scandinavian Journal of Statistics* **24,** 289 – 307.

Wellner, J. and Zhan, Y. (1997). A hybrid algorithm for computation of the nonparametric maximum likelihood estimator from censored data. *Journal of the American Statistical Association* **92,** 945 – 959.

Table 1: Percent bias and coverage to estimate $F(t_0)$, where $t_0 = G^{-1}(0.5)$ is the median of the observation times distribution. Sample sizes of 500 and 1000 observations, with 10% expected failures, were considered. Results are based on 1000 simulations

	$(n = 500)$						$(n = 1000)$					
(ϕ, ψ)	$\hat{b}_n^1$	$\hat{b}_n^2$	$\hat{b}_n^3$	γ_n^2	γ_n^3	$E^{2,3}$	$\hat{b}_n^1$	$\hat{b}_n^2$	$\hat{b}_n^3$	γ_n^2	γ_n^3	$E^{2,3}$
A. Uniform observation times												
$(1, 0.8)$	175.6	-1.2	-2.1	0.944	0.948	93.0	177.2	0.8	0.1	0.947	0.945	92.2
$(1, 0.7)$	264.3	-0.6	-5.4	0.944	0.935	90.0	263.6	-1.8	-3.7	0.950	0.944	88.5
$(0.8, 1)$	-24.5	-5.7	-5.4	0.946	0.951	99.0	-23.5	-4.3	-4.2	0.950	0.942	98.8
$(0.7, 1)$	34.9	-7.0	-6.9	0.947	0.940	99.1	-33.1	-4.4	-4.2	0.949	0.945	98.8
$(0.8, 0.9)$	67.4	-1.3	-3.6	0.946	0.947	92.9	66.7	-2.3	-3.3	0.963	0.941	92.7
$(0.7, 0.8)$	56.4	-3.1	-4.1	0.941	0.945	91.4	56.6	-2.8	-4.2	0.948	0.941	90.1
B. Exponential observation times												
$(1, 0.8)$	252.8	4.4	-1.3	0.952	0.942	92.1	252.0	2.7	-0.5	0.954	0.947	91.2
$(1, 0.7)$	381.8	11.4	1.5	0.948	0.943	89.2	378.4	5.4	-2.4	0.949	0.940	86.9
$(0.8, 1)$	-24.4	-5.5	-5.4	0.942	0.944	99.0	-22.6	-3.3	-3.3	0.948	0.952	98.7
$(0.7, 1)$	-34.1	-5.8	-5.6	0.939	0.935	99.0	-32.3	-3.3	-3.4	0.950	0.952	98.7
$(0.8, 0.9)$	104.7	0.1	-2.6	0.944	0.933	93.2	104.7	-0.2	-2.4	0.948	0.949	91.8
$(0.7, 0.8)$	95.9	2.3	-3.7	0.944	0.942	90.4	94.7	-0.2	-2.6	0.946	0.938	89.8

n = sample size, ϕ = sensitivity, and ψ = specificity
$\hat{b}_n$: Percent bias, $\hat{\gamma}_n$: Nominal coverage
$E^{3,2}$= Average c.i length [3] / Average c.i length [2]
[1] Assuming $\phi = \psi = 1$ (i.e. naive estimator)
[2] Assuming true value (ϕ, ψ)
[3] Assuming true value (ϕ, ψ) and in addition 10% of cohort measured with gold standard

Table 2: Size and power of the proposed test with various samples sizes (200, 500, and 1000), 10 and 20% expected failures, based on 1000 simulations. Naive estimator assumes perfect sensitivity ($\phi = 1$) and perfect specificity ($\psi = 1$). Adjusted estimator assumes the correct sensitivity and specificity

(ϕ, ψ)	p_0	p_1	$n_0 = n_1 = 100$		$n_0 = n_1 = 250$		$n_0 = n_1 = 500$	
			Naive	Adjusted	Naive	Adjusted	Naive	Adjusted
A: Non differential misclassification								
A.1: Under H_0								
(1,0.8)	0.1	0.1	0.052	0.022	0.045	0.030	0.056	0.040
(1,0.7)			0.055	0.021	0.061	0.023	0.050	0.030
(0.8,1)			0.062	0.058	0.054	0.030	0.048	0.038
(0.7,0.9)			0.062	0.026	0.050	0.040	0.048	0.042
(1,0.8)	0.2	0.2	0.056	0.038	0.052	0.040	0.044	0.042
(1,0.7)			0.064	0.042	0.052	0.036	0.048	0.034
(0.8,1)			0.063	0.060	0.062	0.062	0.046	0.042
(0.7,0.8)			0.056	0.040	0.045	0.032	0.047	0.038
A.2: Proportional hazard alternative (Hazard ratio =0.5)								
(1,0.8)	0.1	0.05	0.104	0.048	0.168	0.116	0.276	0.247
(1,0.7)			0.096	0.031	0.132	0.078	0.194	0.148
(0.8,1)			0.222	0.220	0.458	0.456	0.748	0.748
(0.7,0.8)			0.094	0.036	0.162	0.106	0.248	0.216
(1,0.8)	0.2	0.11	0.208	0.157	0.445	0.410	0.717	0.720
(1,0.7)			0.177	0.120	0.334	0.287	0.570	0.549
(0.8,1)			0.385	0.379	0.726	0.723	0.962	0.962
(0.7,0.9)			0.186	0.142	0.357	0.327	0.609	0.588

Table 3: (Continuation) Size and power of the proposed test with various samples sizes (200, 500, and 1000), 10 and 20% expected failures, based on 1000 simulations. Naive estimator assumes perfect sensitivity ($\phi = 1$) and perfect specificity ($\psi = 1$). Adjusted estimator assumes the correct sensitivity and specificity

(ϕ, ψ)	p_0	p_1	$n_0 = n_1 = 100$		$n_0 = n_1 = 250$		$n_0 = n_1 = 500$	
			Naive	Adjusted	Naive	Adjusted	Naive	Adjusted
B: Differential misclassification								
B.1: Under H_0								
(0.9,1) & (0.7,1)	0.1	0.1	0.086	0.052	0.156	0.049	0.224	0.051
(1,0.9) & (1,0.7)			0.820	0.032	0.995	0.034	1.000	0.031
(0.9,0.9) & (0.7,0.7)			0.724	0.024	0.978	0.032	1.000	0.044
B.2: Proportional hazard alternative (Hazard ratio =0.5)								
(0.9,1) & (0.7,1)			0.354	0.191	0.704	0.415	0.948	0.749
(1,0.9) & (1,0.7)	0.1	0.05	0.666	0.021	0.964	0.059	1.000	0.178
(0.9,0.9) & (0.7,0.7)			0.647	0.019	0.954	0.004	1.000	0.022
B.3: Proportional hazard alternative (Hazard ratio =2)								
(0.9,1) & (0.7,1)			0.157	0.352	0.350	0.719	0.575	0.937
(1,0.9) & (1,0.7)	0.1	0.05	0.968	0.204	1.000	0.403	1.000	0.736
(0.9,0.9) & (0.7,0.7)			0.884	0.125	1.000	0.238	1.000	0.401

Table 4: Cox regression coefficients and size of the likelihood ratio test, with samples of 500 observations, 10% expected failures and 1000 simulations

	Model 1: $\Lambda(t) = t\exp(-0.695Z_1)$				**Model 2**: $\Lambda(t) = t\exp(0.405Z_2)$			
(ϕ, ψ)	$\hat{\theta}_n$	$\hat{\theta}_a^{Mean}$	$\hat{\theta}_a^{Median}$	η	$\hat{\theta}_n$	$\hat{\theta}_a^{Mean}$	$\hat{\theta}_a^{Median}$	η
(0.9,0.99)	-0.618	-0.727	-0.700	0.061	0.382	0.421	0.419	0.059
(0.80,0.99)	-0.606	-0.756	-0.687	0.061	0.368	0.417	0.415	0.069
(0.9,0.95)	-0.406	-0.814	-0.738	0.068	0.288	0.424	0.394	0.056
(0.8,0.95)	-0.365	-0.794	-0.701	0.053	0.275	0.434	0.412	0.063

Model 3: $\Lambda(t) = t\exp(-0.695Z_1 + 0.405Z_2)$

(ϕ, ψ)	$\hat{\theta}_{n1}$	$\hat{\theta}_{n2}$	$\hat{\theta}_{1a}^{Mean}$	$\hat{\theta}_{1a}^{Median}$	$\hat{\theta}_{2a}^{Mean}$	$\hat{\theta}_{2a}^{Median}$
(0.9,0.99)	-0.646	0.379	-0.738	-0.704	0.431	0.416
(0.80,0.99)	-0.634	0.385	-0.750	-0.738	0.451	0.439
(0.9,0.95)	-0.451	0.257	-0.787	-0.729	0.430	0.432
(0.8,0.95)	-0.394	0.248	-0.768	-0.707	0.456	0.430

Z_1, Z_2 are independent $Bernoulli(0.5)$ random variables
$\hat{\theta}_n$ = Naive estimate (assuming $\psi = \phi = 1$)
$\hat{\theta}_a$ = Adjusted estimate (assuming true values of ϕ and ψ)
η = size of the likelihood ratio test for testing $H_0 : \theta = \theta_0$

Table 5: Descriptive statistics

	Control (N=933)	Intervention (N=931)	Total (N=1864)
Female*	731 (78.3)	736 (78.5)	1467 (78.7)
Age (years) +	21 [19-25]	22 [19-26]	21 [19-25]
Initial diagnoses			
Gonorrhea	133 (14.3)	132 (14.2)	265 (14.2)
Genital chlamydia	752 (80.6)	752 (80.8)	1504 (80.7)
Both	48 (5.1)	47 (5.0)	95 (5.1)
Events*	122 (13.1)	92 (9.9)	214 (11.5)
Observation time (days) +	87 [77-103]	87 [76-104]	87 [77-103]

Regression Analysis

Factors	Univariate Analysis HR [95 %CI], P-value	Multivariate Analysis HR (P-value)
Intervention	0.744 [0.580,0.956], 0.031	0.565
Gender	1.283 [0.950,1.914], 0.121	1.189
Gender × Intervention	-	1.400 (0.305)

* N(%), + Median [IQR], and HR= Hazard ratio

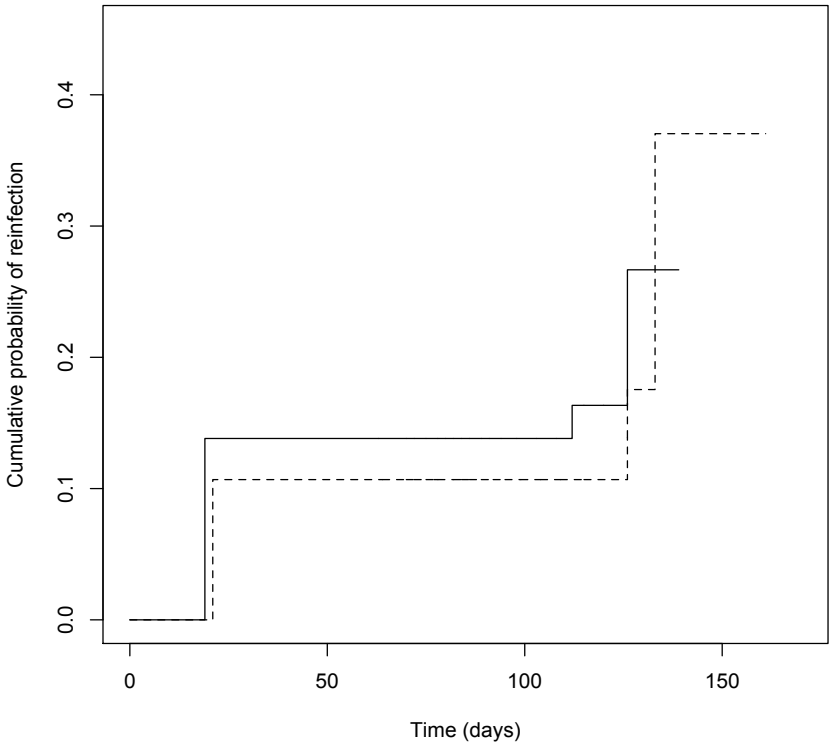

Figure 1: Estimated cumulative probability of reinfection for participants in the placebo (solid line) and intervention (dashed line) arm.