

## Nonparametric and Semiparametric Survival Estimators, and their Implementation, in Two-Stage (Nested) Cohort Studies

Steven D. Mark

Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6130 Executive Plaza South, Room 8036, Executive Blvd., Bethesda, MD, 20852-491. *e-mail*: sm7v@nih.gov.

**SUMMARY.** Frequently in epidemiologic cohort studies, one's primary interest is in estimating the effect of exposures,  $V_i$  on a time-to-event outcome,  $T_i$ , while adjusting for other covariates,  $J_i$ . When the cost of measuring  $V_i$  is disproportionate to the cost of  $J_i$ , it may be inefficient, or infeasible, to ascertain  $V_i$  on everyone. Cost may reflect financial cost, logistical cost, or health risks attendant upon obtaining  $V_i$  measurements from individuals. These considerations have given rise to a number of two-stage sampling schemes where  $J_i$  is observed on all members of a cohort, and  $V_i$  observed only on a subgroup. Common sampling schemes are the case-cohort and nested case-control designs. Various estimating equations, generally focusing on estimating the relative risk parameter in a Cox proportional hazards model, have been proposed. Rather than relative risk, our focus is on survival, or absolute risk. We characterize a class of non and semiparametric cumulative hazard estimators, and estimate survival as a functional of these cumulative hazards. We express differences between estimators within a class in terms of differences in the extent to which subjects with unmeasured  $V_i$  contribute to estimation. We demonstrate how the familiar considerations of identifying what covariates are independent risk factors for disease can impact design and analyses so as to increase the efficiency of estimation. We motivate this work, and demonstrate its novel characteristics with a data analysis, and accompanying simulations, of a two-stage study on *H. pylori* infection and gastric cardia cancer. We provide code written in R language (or S-plus) that implements these estimators.

**KEY WORDS:** absolute risk, auxiliary covariate; case-cohort; cumulative hazard; efficiency; missing at random; nested case-control; population attributable risk, robust estimation; survival; two-stage studies; weighted estimating equations

### 1. Introduction

In epidemiologic cohort studies new exposures, which we call  $V_i$ , frequently become of interest after endpoints have already been recorded at follow-up time  $\tau$ . Two-stage sampling designs are a common strategy for estimating the association of  $V_i$  with outcome in such cohorts. We focus on studies where the outcome is time to some event of interest,  $T_i$ . In the first stage of these studies, one observes a (possibly empty) set of covariates,  $A_i$ , and an outcome  $(X_i, \Delta_i)$  for each  $n$  individuals. As usual,  $X_i = \min(T_i, C_i)$ ,  $C_i$  is a censoring time,  $\Delta_i = I(X_i = T_i)$ , and  $I(\cdot)$  is the indicator function. Throughout this paper we assume censoring is independent and non-informative (see for example, Andersen, Borgan Gill, and Keiding, 1991). Consistent with epidemiologic parlance we call those with  $\Delta_i = 1$  cases, and those with  $\Delta_i = 0$ , controls.  $W_i$  denotes the combined set of outcome and covariate data observed at the end of stage 1 (time  $\tau$ ).

$$W_i = \{X_i, \Delta_i, A_i\} \quad (1)$$

In the second stage of the study, using selection probabilities,  $\pi_o(W_i)$ , that depend only on  $W_i$ , a sub-sample of individuals is chosen for measurement of  $V_i$ . The motivation for sub-sampling is that  $V_i$ , which we subsequently refer to as the *exposures*, are, in some sense, expensive, or difficult, to measure. Since the occurrence of cases is rare compared to that of controls, and case counts are the main determinants of the variance of the estimators, sampling rates are generally higher for cases than for controls. We define  $R_i = 1$  if  $V_i$  is known for individual  $i$ ;  $R_i = 0$  otherwise. To control confounding, an investigator generally estimates the effect of  $V_i$  conditional on a set of *adjusting covariates*,  $J_i$ ,  $J_i \subseteq A_i$ . We call variables in  $A_i$  that are not in  $J_i$ , *auxiliary variables*, and denote them  $\Lambda_i^{aux}$ .

When the outcome is time-to-event, the most common two-stage designs are the case-cohort (Prentice, 1986; Self and Prentice, 1988) and nested case-control designs (Lidell, McDonald, Thomas, 1997; Borgan, Goldstein, and Langholz, 1995). The primary focus of these designs has been estimating relative risks ( $rr$ ) associated with covariates  $Z_i = \{V_i, J_i\}$ , when hazards are specified by a Cox proportional hazards model (CPH) such as (4). We recently reviewed these approaches and showed that the estimators and their variances can be written as a single set of estimating equations (Mark and Katki, 2001). In contrast, rather than estimating

$\beta_o$ , the focus in this paper is on the estimation of the conditional survivals,  $S(\tau|z)$ , where  $z \in \mathcal{Z}$  (the support of  $Z_i$ ). In Appendix A.4 we provide estimators for functionals of the conditional survivals such as standardized survivals, standardized risk differences, and population attributable risks.

The motivation for this work arises from two-stage studies we have conducted on a cohort in China with epidemic rates of gastric cardia stomach cancer (GCC). This cohort was selected from a well defined geographic population, and estimates of survival, or absolute risk, are of public health importance (Mark, Qiao, Dawsey, et al., 2000). To illustrate the underlying issues, we use data from our study on the association of *H.Pylori* (Hp) infection with incident GCC (Limburg PJ, Wang CQ, Mark SD, et al., 2000). In that study,  $V_i$  was the measurement of serum antibodies to Hp:  $V_i=1$  if a subject had antibodies to Hp.  $A_i$  contained such information as age, sex, height, and weight. Since age was the only significant risk factor in  $A_i$ , in the analysis we present,  $J_i$  is an indicator variable, with  $J_i = 1$  if a subject's age is greater than the median cohort age. We document the performance of estimators using simulations based on the structure of a current study using the endpoints accrued over a subsequent ten years.

Though our inferential focus is survival, the mathematical results we present are on the cumulative hazard scale,  $\Lambda(t; z)$ ,

$$\Lambda(t; z) = \int_0^t \lambda(u|z^\dagger) du \quad 0 \leq t \leq \tau \quad (2)$$

We obtain survival estimators through the identity

$$S(t|z) = \exp - (\Lambda(t; z)) \quad (3)$$

In 1994 Robins, Rotnitzky and Zhao (1994) (henceforth called RRZ), described the class of all two-stage estimators in terms of weighted estimating equations, and derived the mathematical form of the efficient member of the class. They focused on conditional mean models. Applying their results to time-to-event data, we describe the class of nonparametric and semiparametric cumulative hazard estimators. In nonparametric estimation no assumptions are made regarding the relationship between hazards at different levels of  $z$ . For the semiparametric model we assume the hazards are related by the Cox proportional hazards (CPH) model

$$\lambda(u|Z_i) = \lambda_o(u) \exp (\beta_1^T V_i + \beta_2^T J_i); \quad (4)$$

where  $\{V_i, J_i\}$  and  $\{\beta_1, \beta_2\}$  are conformable  $p \times 1$  vectors of covariates and parameters. For simplicity, we assume that the  $Z_i$  are time invariant, and that, as expressed in (4), there is no  $V$  by  $J$  interaction. We refer to  $\lambda_o(u)$  as the baseline hazard. Since our interest is in contrasting survivals of groups of individuals, in the body of the paper we assume  $Z_i$  has finite support of dimension  $k^*$ . In the Hp data  $k^* = 4$ . In appendix D we give results on estimation in a completely general support space.

There are several practical advantages to the RRZ formulation. Both the case-cohort (CCH), and nested case-control (NCC) designs specify that cases be sampled with probability one. When  $V_i$  is not measured on all cases, the cumulative hazard estimators given in those proposals are biased (Mark and Katki, 2001). Epidemiology studies commonly require exposure measurements which are expensive and consume limited specimens. Consequently designs with fractional cases-sampling have become increasingly frequent and attractive (Mark and Katki, 2001). In the Hp study, due to uncertainties with regard to the direction of the association, and the prevalence of Hp infection, as well as a reluctance to use up the small quantities of available serum, we sampled approximately 25% of available GCC cases. The estimating equations we describe are weighted by the inverse of the sampling probability, and accommodate any non-zero sampling rate. Even when the intent of an investigator is to measure  $V$  on all the cases, vagaries beyond investigator control seldom, if ever, permit complete ascertainment (Mark et al, 2000; Mark and Katki, 2001). We describe the additional assumptions required for estimation when there is unplanned missingness in section 6.

Another feature distinguishing RRZ from CCH and NCC estimators, is that in RRZ estimators, individuals with unobserved  $V_i$  contribute to estimation. In this paper we emphasize that the distinction between estimators within the RRZ class, and hence the differences in efficiency, are entirely due to variation in the extent to which information from subjects with  $R_i = 0$  is utilized. We derive expressions for the efficient nonparametric, and the restricted-class efficient (defined in section 4) semiparametric estimators, in terms of aspects of the probability distribution of the data familiar to epidemiologists. This formulation has clear implications for the design and analysis of two-stage studies.

Finally, in this paper we emphasize a particular approach to estimation, which we call  $\hat{\pi}$ -estimation. Formulation in terms of  $\hat{\pi}$ -estimation provides a geometric representation, as well as a practical means of implementing, efficiency consideration. R and S-plus code that implements these  $\hat{\pi}$ -estimators is provided in Appendix F.

## 2. Full-Data Estimators and Influence Functions

We refer to studies in which  $V_i$  is observed for all  $n$  individuals as *full data studies*, and use  $H_i = \{W_i, V_i\}$  to denote the fully observed data. In a sense made specific in section 4, RRZ proved that that all two-stage estimators and their corresponding influence functions can be expressed as *weighted versions with offset* of their full data counterparts. In this section we describe the full data estimators and influence functions for the nonparametric cumulative hazard, and for the semiparametric estimators of  $\beta_o$  and the baseline cumulative hazard (5).

$$\Lambda_o(\tau, \beta_o) = \int_0^\tau \lambda_o(u) du \tag{5}$$

For the semiparametric model, the cumulative hazard at any covariate level  $z$  is  $\Lambda(\tau; \beta_o z) = \Lambda_o(\tau, \beta) \exp \beta_o^T z$ , and is estimated in the obvious fashion. Its distribution is derived by the delta method, such as in Anderson et al. (1991). To indicate the  $k^* \times 1$  vector of cumulative hazards, we drop  $z$  from the arguments and write  $\Lambda(\tau)$ , or  $\Lambda(\tau, \beta_o)$ . For concreteness we generally use  $\tau$  as the time argument in the cumulative hazards or survival. In section 4 we provide corresponding results for two-stage estimators.

In full data studies, the Nelson -Aalen estimator,  $\hat{\Lambda}(\tau, z)$ , is the efficient nonparametric estimator of (2) (Anderson et al., 1991). The partial likelihood estimator,  $\hat{\beta}$ , and the Breslow estimator,  $\hat{\Lambda}_o(\tau, \hat{\beta})$  are the semiparametric efficient estimators of  $\beta_o$  (4) and the baseline cumulative hazard (5) (Anderson et al, 1991). Using standard counting process notation, we denote the event counting process  $N_i(u)$ , ( $N_i(u) = 1$  iff  $T_i \leq u$ , and  $T_i \leq C_i$ ), and the at risk process,  $Y_i(u)$ , ( $Y_i(u) = 1$ , iff  $(C_i \wedge T_i) \leq u$ ). For individual  $i$  the hazard of  $T_i$  conditional on  $Z_i$  is  $\lambda_i(u|Z_i) = Y_i(u) \times \lambda(u|Z_i)$ . We assume the  $\lambda(u|z)$  are non-negative, and the  $\Lambda(\tau; z)$  are finite. In the context of nonparametric estimation the above should be regarded as a multivariate counting process of dimension  $k^*$ . That is, we estimate the  $k^*$  within-stratum cumulative hazards,  $\Lambda_h(\tau)$ , where, for instance, the  $h$ 'th row of  $N_i(u)$  is  $N_{ih}(u) = 1$ , iff  $I(Z_i = h), T_i \leq u$ , and  $T_i \leq C_i$ . As is standard we define,  $S^0(u) = \sum_{j=1}^n Y_j(u)$ ;  $S^0(u, \hat{\beta}) = \sum_{i=1}^n Y_i(u) \exp(\hat{\beta} Z_i)$ ; and  $S^1(u, \hat{\beta}) = \sum_{i=1}^n Y_i(u) Z_i \exp(\hat{\beta} Z_i)$ .

Under the usual regularity conditions (Anderson et al., 1991),  $n^{-1} S^j(u, \cdot) \xrightarrow{lim p} E[S^j(u, \cdot)] = s^j(u, \cdot)$  for all three processes.  $M_i(u)$  denotes the counting process martingale,  $M_i(u) = N_i(u) - \Lambda_i(u)$ .

The  $k^* \times 1$  full data Nelson-Aalen estimator of  $\Lambda(\tau)$  is (Anderson

$$\hat{\Lambda}(\tau) = \sum_{i=1}^n \int_0^\tau S^0(u)^{-1} dN_i(u) \tag{6}$$

Anderson et al. (1991) give the influence function expansion of  $\hat{\Lambda}(\tau)$  as

$$n^{\frac{1}{2}} \left( \hat{\Lambda}(\tau) - \Lambda(\tau) \right) = n^{-\frac{1}{2}} \sum_{i=1}^n D_i^{F1} + o_p(1) \tag{7}$$

$$D_i^{F1} = \int_0^\tau [s^0(u)]^{-1} dM_i(u) \tag{8}$$

Newey (1990) showed that all nonparametric estimators have identical influence functions, and hence, are asymptotically equivalent. Thus (8) is the influence function for any nonparametric estimator of  $\Lambda(\tau)$ .

The class of full data estimators for the  $\beta_o$  in CPH model (4) (RRZ, 1994) are the  $\hat{\beta}(h)$ 's that solve

$$\sum_{i=1}^n \int_0^\tau \left\{ h(Z_i, X_i) - S^1(s, \beta, h) S^0(s, \beta)^{-1} \right\} dN_i(s) = 0 \tag{9}$$

The choice of the function  $h(Z_i, X_i)$  determines the efficiency of the estimator. For full data, the semiparametric efficiency bound is achieved by the partial likelihood estimator with  $h(Z_i, X_i) = Z_i$ . The full data influence function for the partial likelihood estimator is  $D_i^{F2}$

$$D_i^{F2} = i^{-1} \int_0^\tau \left\{ Z_i - e(u, \beta_o) \right\} dM_i(u) \tag{10}$$

where  $e(u, \beta_o) = s^1(u, \beta_o) s^0(u, \beta_o)^{-1}$ , and  $i = E[D_i^{F2} D_i^{F2'}]$ , the usual partial likelihood information. As in (7), the estimator  $\hat{\beta}$  can be expressed as the sum of its iid influence functions.

The Breslow estimator of the baseline cumulative hazard,  $\Lambda_o(\tau, \beta)$ , is given by

$$\hat{\Lambda}_o(\tau, \hat{\beta}) = \sum_{i=1}^n \int_0^\tau [S^0(u, \hat{\beta})]^{-1} dN_i(u) \tag{11}$$

To obtain the influence function for (11) we write

$$\widehat{\Lambda}_o(\tau, \widehat{\beta}) - \Lambda_o(\tau, \beta_o) = \left\{ \widehat{\Lambda}_o(\tau, \widehat{\beta}) - \widehat{\Lambda}_o(\tau, \beta_o) \right\} + \left\{ \widehat{\Lambda}_o(\tau, \beta_o) - \Lambda_o(\tau, \beta_o) \right\} \tag{12}$$

Using a Taylor series expansion of  $\widehat{\beta}$  around  $\beta_o$  as in Theorem VII.2.3 Anderson et al. (1991), we express the first term in the right hand side of (12) as

$$(\widehat{\beta} - \beta_o)' \int_0^\tau e(u, \beta_o) \lambda_o(u) du + o_p(1)$$

Then replacing estimators in (12) with their influence functions, we have

$$n^{\frac{1}{2}} \left\{ \widehat{\Lambda}_o(\tau, \widehat{\beta}) - \Lambda_o(\tau, \beta_o) \right\} = n^{-\frac{1}{2}} \sum D_i^{F3} + o_p(1) \tag{13}$$

$$D_i^{F3} = \int_0^\tau [s^0(u, \beta_o)]^{-1} dM_i(u) - D_i^{F2'} \int_0^\tau e(u, \beta_o) d\Lambda_o(u, \beta_o)$$

We refer to the  $D_i^{Fb}, b \in \{1, 2, 3\}$ , as the *full data influence functions*. Like all influence functions, they are iid and have expectation 0. Hence the asymptotic variance of each estimator is  $E \left[ D_i^{Fb} D_i^{Fb'} \right]$ .

Though we explicitly present results for a non-stratified CPH model, the results of a stratified model, with strata generated from a discretization of  $A_i$ , can be described using the multivariate counting process.

### 3. Stage-Two Sampling Restrictions

For most of the paper we assume that conditional on  $W_i$ , selection of individuals for measurement of  $V_i$  is independent with known, non-zero, probabilities,  $\pi_o(W_i)$  that do not depend on  $V_i$ . That is

$$\pi(W_i) = Pr(R_i = 1 | W_i, V_i) = Pr(R_i = 1 | W_i) \tag{14}$$

In the usual parlance of missing data, restriction (14) is consistent with  $V_i$  being missing at random (MAR) (Rubin, 1976). As we frequently do for random variables, we drop the explicit argument of a function, and use the subscript  $i$  to indicate that it is a random variable. Thus we write  $\pi_{i,o}$ , where  $\pi_{i,o} \equiv \pi_o(W_i)$ . At the end of section 8 we extend the results to dependent sampling, and to missingness that is not entirely under investigator control.

Without loss of generality we specify the known sampling probabilities using the logistic model

$$\text{logit } \pi_o(W_i) = \psi_o' h(W_i) \tag{15}$$

Here  $\psi_o$  and  $h(W_i)$  are known, conformable, finite dimensional vectors of parameters and random variables, respectively. Clearly neither the parameterization nor the dimension of equation (15) are unique. For instance, if  $A_i$  contains only information on sex, and stage-two sampling depends only on case status, then two correctly specified models for (15) would be

$$\text{logit } \pi_o(W_i) = \psi_{o1} I(\Delta_i = 1) + \psi_{o2} I(\Delta_i = 0) \tag{16}$$

$$\text{logit } \pi_o(W_i) = \psi_{o1} I(\Delta_i = 1) + \psi_{o2} I(\Delta_i = 0) + \psi_{o3} I(\text{male}) + \psi_{o4} I(\text{female}) \tag{17}$$

Here  $\psi_{o1} = \text{logit } Pr(R_i = 1 | \Delta_i = 1)$ ;  $\psi_{o2} = \text{logit } Pr(R_i = 1 | \Delta_i = 0)$ , and  $\psi_{o3} = \psi_{o4} = 0$ .

We define  $W_i^R$  to be the smallest set of linearly independent vectors such that (15) is true where size refers to the dimension of the column space spanned by the  $h(W_i)$ . In our example, the dimension of  $W_i^R$  is two. Correctly specified models are those with covariates  $W^l$  such that

$$W_i^l \geq W_i^R. \tag{18}$$

We consider models with equivalent spans to be identical, and restrict ourselves to covariate spaces where the  $W_i^l$  are linearly independent.

We denote the scores from any logistic model with covariates  $W_i^l$  as  $S_i^l$ ,

$$S_i^l = (R_i - \pi_{i,o}) W_i^l \tag{19}$$

### 4 Two-Stage Estimators and Influence Functions

The two-stage risk set estimators,  $\widetilde{S}^j(u, \cdot)$ , are inverse probability weighted versions of the full data estimators. For instance,  $\widetilde{S}_h^0(u) = \sum_{j=1}^n \pi_{i,o}^{-1} R_i Y_{i,h}(u)$ . Like their full data counterparts, their averages converge in probability to  $s^j(u, \cdot)$  (Pugh, 1993; RRZ, 1994).

RRZ prove that two-stage estimators, and their influence functions, can be expressed as weighted versions of the full data quantities with an "offset". Applying these results to nonparametric estimation establishes that all two-stage estimators of  $\Lambda(\tau)$  are asymptotically equivalent to a member in the class of estimators,  $\widetilde{\Lambda}(\tau, g_1)$ , defined as

$$\tilde{\Lambda}(\tau, g_1) = \sum_{i=1}^n \left\{ \int_0^\tau R_i \pi_{i,o}^{-1} \left( \tilde{S}^0(u) \right)^{-1} dN_i(u) - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_1(W_i) \right\} \tag{20}$$

Here  $g_1(W_i)$  is any  $k^* \times 1$  vector of non-stochastic functions of  $W_i$  specified by the investigator. The corresponding influence functions are

$$D_i^1(g_1) = R_i \pi_i^{-1} D_i^{F1} - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_1(W_i) \tag{21}$$

Note that here, unlike the full data case where we have a single estimating equation (6) and influence function (8), there are a class of estimators and influence functions characterized by the "offset"  $\pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_1(W_i)$ .

The class of two-stage semiparametric estimators of  $\beta_o$  are characterized by an  $h(\cdot)$  function as well as the  $p \times 1$  offset term,  $\pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_2(W_i)$  (RRZ,1994). Efficiency in estimation depends on the choice of both the  $h(\cdot)$  and  $g_2(\cdot)$  functions. The optimal function  $h(\cdot)$  is a non-closed form integral equation that is a function of infinite dimensional parts of the survival and covariate distributions (RRZ, 1994). For reasons of practicality, we therefore follow a general recommendation of RRZ and restrict our estimators to the subclass that use the efficient full data  $h(\cdot)$  function,  $h(Z_i, X_i) = Z_i$ . This subclass includes the CCH and NCC estimators. Hence, we consider estimators  $\tilde{\beta}(g_2)$  that solve

$$\sum_{i=1}^n \left( \int_0^\tau R_i \pi_i^{-1} \times \left\{ Z_i - \tilde{S}^1(s, \beta) \tilde{S}^0(s, \beta)^{-1} \right\} dN_i(s) - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_2(W_i) \right) = 0 \tag{22}$$

Due to this restriction, we refer to efficiency results for estimators of  $\beta_o$ , and  $\Lambda_o(s, \beta_o)$  as restricted-class efficient estimators (RC-efficient). The influence functions for  $\tilde{\beta}(g_2)$  are

$$D_i^2(g_2) = \pi_{i,o}^{-1} R_i D_i^{F2}(\beta) - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_2(W_i) \tag{23}$$

The procedure for estimating the baseline cumulative hazard (5) is analogous to the full data case.

First estimate  $\tilde{\beta}(g_2)$ , then, with  $g_3^*(W_i)$  any scalar function of  $W_i$ ,

$$\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*) = \sum_{i=1}^n \pi_{i,o}^{-1} \left\{ R_i \int_0^\tau [\tilde{S}^0(u, \tilde{\beta}(g_2))]^{-1} dN_i(u) - (R_i - \pi_{i,o}) g_3^*(W_i) \right\}; \tag{24}$$

Using an identical Taylor series expansion as above, the influence functions for (24) are

$$D_i^3(g_3) = \pi_{i,o}^{-1} R_i D_i^{F3} - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_3(W_i); \quad g_{i,3} = g_{i,3}^* - g_{i,2} \int_0^\tau e(u, \beta_o) d\Lambda_o(u, \beta_o) \tag{25}$$

As for the full data case, the asymptotic variances of the estimators are  $E \left[ D_i^b D_i^{b'} \right]$ .

Letting  $\tilde{\Lambda}(\tau, \cdot)$  denote any non or semiparametric estimator of (3), we estimate  $S(\tau|vj)$  by substituting  $\tilde{\Lambda}(\tau, \cdot)$  for  $\Lambda(\tau)$  in (3). The asymptotic variances are obtained by the delta method. Consistent estimators of  $E \left[ D_i^{Fb} D_i^{Fb'} \right]$ , and the variances of the  $\tilde{S}(\tau, \cdot |vj)$ , are given in appendix A.

### 5. The Efficient $g_b(\cdot)$ for Estimators of $\Lambda(s)$ and $\beta_o$

We define simple true- $\pi$  estimators (STP) as those in which  $g_b = 0$  in 20,22,24, and write the STP influence functions as,  $D_i^b(\pi_o) \equiv \pi_{i,o}^{-1} R_i D_i^{Fb}$ . We can then express the  $D_i^b(g_b)$  as

$$D_i^b(g_b) = D_i^b(\pi_o) - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_{i,b} \tag{26}$$

From 20,22,and 24, it is apparent that differences between estimators in each class are entirely due to differences in the  $g_b$ . Thus, finding the minimum variance estimator is equivalent to finding the  $g_b$  that minimizes  $E [ D_i^b(g_b) D_i^{bT}(g_b) ]$ . We call such a  $g_b$ , the efficient, or for the semiparametric models RC-efficient,  $g_b$ , and denote it by  $g_b^{eff}$ . For  $b \in \{1, 2\}$ , direct application of proposition 2.3 of RRZ establishes that  $g_{i,b}^{eff} = E [ D_i^{Fb} | W_i ]$ . For estimators of  $\Lambda_o(s, \beta)$ , which are a function of  $g_2$  and  $g_3^*$ , we use RRZ 2.3 and show (Appendix B) that the minimum variance is obtained with  $g_2 = 0$ ,  $g_3^* = E [ D_i^{Fb} | W_i ]$ , and that  $g_3^{eff} = E \left[ D_i^{F3} | W \right]$ .

RRZ prove that an equivalent representation of the influence functions in (26) is

$$D_i^b(W^l) = D_i^b(\pi_o) - q^b S_i^l \tag{27}$$

Here  $q^b$  is any conformable matrix of constants, and  $S_i^l$  are scores (19) from correctly specified logistic models (15). In appendix C we use (27) to provide an alternative derivation for the efficiency results of RRZ 2.3. The proof relies on the following two characteristics of population least squares regression that are fundamental to understanding the  $\hat{\pi}$ -estimating procedures, their efficiency properties, and their method of implementation: 1) for any given set of scores,  $S_i^l$ , the variance of  $D_i^b(W^l)$  is minimized when  $q^b$  is the projection operator,  $P^{bl}$ , of  $D_i^b(\pi_{i,o})$  on  $S_i^l$ .

$$P^{bl} = E [ D_i^b(\pi_o) S_i^{l'} ] E [ S_i^l S_i^{l'} ]^{-1} \tag{28}$$

2) Since  $D_i^b(\pi_{i,o}) - P^{bl}S_i^l$  is the residual from a projection, the variance is non-decreasing in the dimension of  $W_i^l$ . In appendix C we show that the minimum variance is reached when

$$P^{bl}S_i^l = \pi_{i,o}^{-1} (R_i - \pi_{i,o}) E \left[ D_i^{Fb} | W_i \right] \tag{29}$$

and that (29) (Appendix C, Result 2) is true for logistic model (30)

$$\text{logit } \pi_o(W_i) = \psi_1' h(W_i) + \psi_2' W_{i,b}^{eff}; \quad W_{i,b}^{eff} = \pi_{i,o}^{-1} E \left[ D_i^{Fb} | W_i \right] \tag{30}$$

### 6. $\hat{\pi}$ -Estimators

We define  $\hat{\pi}$ -estimators to be the solution to estimating equations 20,22, and 24 when  $g_{i,b} = 0$  and the known sampling probabilities,  $\pi_{i,o}$ , are replaced with predicted sampling probabilities,  $\hat{\pi}_i(W^l)$ . Specifically, the  $\hat{\pi}_i(W^l)$  are formed by replacing  $\psi_o$  in (15) with maximum likelihood estimates,  $\hat{\psi}$ . RRZ (proposition 6.1) show that  $\hat{\pi}$ -estimators are consistent, asymptotically normal, with influence function

$$D_i^b(\hat{\pi}(W^l)) = D_i^b(\pi_{i,o}) - P^{bl}S_i^l \tag{31}$$

It immediately follows that the variance of any  $\hat{\pi}(W^l)$ -estimator is less than or equal to the variance of the STP estimator; and, for  $W^m > W^l$ , the variance of the  $\hat{\pi}(W_i^m)$  estimator is less than or equal to the variance of the  $\hat{\pi}(W_i^l)$  estimator. In Result 1 appendix C, we show that for  $\hat{\pi}$ -estimators based on a model saturated in  $W^f$ , such as model (17),  $P^{bf}S_i^f = \pi_{i,o}^{-1}(R_i - \pi_{i,o})E \left[ D_i^{Fb} | W_i^f \right]$ . In Appendix F we provide code for implementing  $\hat{\pi}$ -estimators using any logistic model (15).

One feature of  $\hat{\pi}$ -estimation is that it is the "natural" estimating procedure when the requirements that sampling is independent and with known probabilities are relaxed. In general, the dependent sampling we consider is characterized as follows: partition the observed  $W_i$  into a finite number of strata; select a fixed number of cases and controls from each stratum. If we let  $W_i^f$  be the saturated column space of indicator variables generated by that partition, then we can use any  $\hat{\pi}$ -estimator with  $W^l \geq W^f$  (RRZ, lemma 6.2). Such dependent sampling commonly occurs. For example, in the Hp study we sampled a fixed number of cases and controls. NCC risk set sampling is by design dependent. We review the definition of NCC sampling and provide appropriate  $\hat{\pi}$ -estimators in Appendix E. We have so far assumed that  $\pi_{i,o}$ , or equivalently, the  $\psi_o$  in specified logistic models (15, 18) are known. If rather than knowing  $\psi_o$ , we only know there is a  $\psi^*$ , such that  $\text{logit } \pi_{i,o} = \psi^*W_i^l$ , then the estimator  $\hat{\pi}(W_i^l)$  has influence function given by (31) (RRZ, proposition 6.2).

### 7. Analyses of the Hp Data Using the $\hat{\pi}(\Delta, J)$ -estimator

Though Hp infection is a well established risk factor for gastric cancers arising outside of the cardia of the stomach (Helicobacter and Cancer Collaborative Group, 2001), the association with gastric cancers that arise in the cardia region (the proximal 2-3 centimeters of the stomach) is less established. Prior to our study, only a few small studies (case sizes ranging from 4 to 12), examined the Hp-GCC association. The consensus from these studies (Helicobacter and Cancer Collaborative Group, 2001; Dawsey, Mark, Taylor, et al., 2002), all conducted on Western populations, was that Hp was "protective" for GCC, with  $rr \approx 0.5$ . Various mechanistic hypothesis have been advanced to account for the opposite association of Hp on GNC and GCC (Blaser, 1999).

In our study (Limburg et al. 2001) we sampled approximately 25% of GCC cases (100 cases) and 7% of controls (200 controls) that occurred in the cohort of 30,000 by 5.25 years of follow-up. We found an Hp prevalence ( $Hp^+$ ) of approximately 65%, and a  $rr$  of approximately two for  $Hp^+$  individuals. The only other major independent risk factor for GCC in this population was age: age greater than the cohort median age increased GCC risk by a factor of 3.5.

Table 1 contains estimates of covariate specific survivals at  $\tau = 5.25$  years based on the CPH model (4) with  $V$  (Hp) and  $J$  (age) indicator variables as defined in the introduction. We used the  $\hat{\pi}$ -estimator based on logistic model (17). Throughout this paper we denote this estimator as  $\hat{\pi}(\Delta, J)$ . At each level of age, the  $Hp^+$  group had lower survivals than the  $Hp^-$  group. Within levels of Hp exposure, survival was higher in the younger group. Using the population age distribution for standardization (see A.4 for definitions and formulae), we estimated that  $Hp^+$  individuals had 1.8% more cases (95% CI, 0.02-2.15) of GCC than the  $Hp^-$  individuals in the 5.25 years of follow-up. (*INSERT TABLE 1 HERE*)

## 8. Implications of Efficiency for Study Design and Analysis

### 8.1 The general case

The optimal  $g(\cdot)$  function,  $E\left[D_i^{Fb}\middle|W\right]$ , is a function of unknown parameters. RRZ proposition 2.4 established that  $g_b^{eff}$  can be replaced by a consistent estimator,  $\widehat{g}_b^{eff}$ , without changing the asymptotic distribution of the estimator. That is, an estimator using  $\widehat{g}_b^{eff}$  achieves the nonparametric efficiency, or semiparametric RC-efficiency, bound. When  $g_b^{eff}$  can be consistently estimated from the data and model assumptions, we say the efficient estimator is identified. If not, then the variance of the efficient influence function represents an unknown lower bound that no estimator is guaranteed to achieve. It is immediately clear that unless  $X_i$  is a deterministic function of  $\{\Delta_i, A_i\}$ ,  $E\left[D_i^{Fb}\middle|W\right] \neq E\left[D_i^{Fb}\middle|\Delta_i, A_i\right]$ , and efficient estimation requires  $X_i$  in the conditioning event. In the remainder of this section we approach the task of conditioning on  $X_i$ , by re-expressing  $g_b^{eff}$  in terms of relative risks, survivals, and covariate distributions. We discuss conditions under which each of these can be consistently estimated, and examine the implications for study design and analysis.

We re-express  $g_{i,b}^{eff}$  as

$$g_{i,b}^{eff} = EE\left[D_i^{Fb}\middle|W_i, V_i\right] = \int_{\mathcal{V}} D_i^{Fb}(W_i, v) Pr(v|W_i) dv \quad (32)$$

In the design stage, a crucial consideration is what, if any, auxiliary variables should be measured. From (32) it is clear that for  $\Lambda_i^{aux}$  to be optimal, it is sufficient that for any larger set,  $\Lambda_i^{aux+} > \Lambda_i^{aux}$ ,

$$Pr(v|X_i, \Delta_i, J_i, \Lambda_i^{aux}) = Pr(v|X_i, \Delta_i, J_i, \Lambda_i^{aux+}) \quad (33)$$

That is, we should collect all auxiliary information which provides additional knowledge about the distribution of the incompletely measured covariates  $V_i$  at any time on study. To further examine the determinants of  $Pr(v|W_i)$ , we reparameterize in terms of the time dependent exposure odds

$$K_{i,v^\dagger} \equiv K_{v^\dagger}(W_i) = Pr(V_i = v^\dagger | \Delta_i, X_i, A_i) / Pr(V_i = v^1 | \Delta_i, X_i, A_i) \quad (34)$$

where  $v^1$  is some chosen reference level in  $\mathcal{V}$ , and  $v^\dagger \in \mathcal{V}$ . Using Bayes' theorem and a non-informative censoring assumption we show (Appendix D) that

$$K_{i,v^\dagger} = rr(X_i|v^\dagger, A_i)^{\Delta_i} \times S(X_i|v^\dagger, A_i) / S(X_i|v^1, A_i) \times Pr(v^\dagger|A_i) / Pr(v^1|A_i) \quad (35)$$

Here  $rr(X_i|v^\dagger, A_i)$  and  $S(X_i|v, A_i)$  are the relative risks and survival probabilities at  $X_i$  conditional on  $\{V_i, A_i\}$ , rather than  $\{V_i, J_i\}$ . By (35), (33) is true if, for all times  $u$

$$S(u|V_i, J_i, \Lambda_i^{aux}) = S(u|V_i, J_i, \Lambda_i^{aux+}) \quad (36)$$

and

$$Pr(V_i|J_i, \Lambda_i^{aux}) = Pr(V_i|J_i, \Lambda_i^{aux+}) \quad (37)$$

Epidemiologists refer to (36) as  $\Lambda_i^{aux}$  containing all *independent predictors of outcome*; and (37) as  $\Lambda_i^{aux}$  containing all *independent predictors of exposure*.

The requirements for efficient analysis are conceptually and mathematically equivalent to those in the design stage. That is, to estimate  $g_b^{eff}$ , we need only include in the conditioning event that subset of  $\Lambda^{aux}$  that contains the independent predictors of outcome and exposure. Though for any given  $\Lambda_i^{aux}$  it is impossible to know with certainty whether (36) or (37) are true, these are the exact considerations required to control confounding. Consequently, in the analysis stage epidemiologists generally try to choose  $J_i$  as the subset of  $A_i$  such that (36) and (37) are "approximately" true when  $\Lambda^{aux}$  is removed from the conditioning event. If successful,  $J_i$  contains all the independent risk factors of outcome, and (35) becomes

$$K_{i,v^\dagger} = rr(X_i|v^\dagger, J_i)^{\Delta_i} \times S(X_i|v^\dagger, J_i) / S(X_i|v^1, J_i) \times Pr(v^\dagger|A_i) / Pr(v^1|A_i) \quad (38)$$

## 8.2 Efficiency when $J$ contains all the independent risk factors

In this section we assume that there is no confounding, in particular that  $J_i$  contains all independent risk factors such that (38) is true. From (32) it is clear that we can estimate  $g_{i,b}^{eff}$ , if we can estimate each of the terms in  $K_{i,v^\dagger}$ . For both the non and semiparametric failure time models, the second and third terms can be estimated by  $\widetilde{S}(X_i|v^\dagger, J_i)$  and  $\widehat{P}(v^\dagger|A_i)$ , where  $\widehat{P}(v^\dagger|A_i)$  is the empirical average of  $V$  within levels of  $A$ . For the semiparametric model,  $rr(u|Z_i)$  can be estimated by  $r\widetilde{r}(u|Z_i) = exp\widetilde{\beta}^T Z_i$ . Here the  $\widetilde{S}(X_i|v^\dagger, J_i)$  and  $\widetilde{\beta}$  come from estimates based on any  $g_b$ . Hence the semiparametric RC-efficient estimators of  $\beta_o$  and  $\Lambda_o(\tau, \beta)$   $exp\beta_o^T Z$  are identified. In contrast, the nonparametric model provides no obvious estimator of  $rr(u|Z_i)$ . If  $k^*$  were small, and the number of cases large, one could theoretically use kernel smooths to estimate hazards, and hence  $rr$ 's. We do not explore this possibility further. Instead, in section 9 we propose several *locally*

efficient estimators (LE-estimators). LE-estimators approximate  $g_b^{eff}$  by making assumptions about  $rr(u|Z_i)$ . We denote the resultant approximations by  $\hat{g}_b^{2eff}$ . If the assumptions about the  $rr$ 's are correct, then  $\hat{g}_b^{2eff} \xrightarrow{lim P} g_b^{eff}$ , and the LE-estimators are efficient. Regardless of the truth of the assumptions, the proposed LE-estimators are consistent.

### 9. Simulations of STP, $\hat{\pi}$ , RC-efficient and Locally Efficient Estimators

All simulations are based on the following covariate distribution:  $Pr(J1) = 0.5$ ,  $Pr(V1) = 0.65$ ;  $Pr(V1|J1) = 0.85$ .  $T_i$  was specified by a CPH model with exponential baseline hazard. The magnitudes for the baseline hazard, and the exponential hazard for the independent censoring times, were chosen to produce approximately 1000 expected cases in a cohort of size  $n=6600$  by time  $\tau$ . This is the approximate number of cases that have occurred through the latest endpoint assessment,  $\tau=15$  years. For the semiparametric models we simulated under the two covariate CPH model (4) with  $\beta_1 = \ln 2$  ( $rr_v = 2$ ),  $\beta_2 = \ln 3$  ( $rr_j = 3$ ), and estimated  $S(\tau|vj)$ . For the nonparametric simulations the data were generated by a one-covariate CPH model with  $\beta_1 = \ln 2$  ( $rr_v = 2$ ); we estimated  $S(\tau|v)$ . Stage 2 sampling was always binomial, depending only on case status (16). Fifteen per cent of controls and 25% of cases were sampled, with a resultant control-to-case ratio of approximately 3:1. Each of the simulation results represents the average of 2000 realizations. Since, as evident from the tables, all of the survival estimators (and estimators of  $\beta_o$ , data not shown) were unbiased and had confidence intervals that covered near the stated rates, we focus the discussion on *relative efficiency (RE)*. RE is defined as the ratio (times 100) of the variance of a given estimator to the variance of the STP estimator. The smaller the RE, the greater the efficiency.

Table 2 contrasts the STP, RC-efficient, and  $\hat{\pi}(\Delta, J)$  semiparametric estimators of  $S(\tau|v0j0)$  and  $S(\tau|v1j1)$ . Both the RC-efficient and the  $\hat{\pi}(\Delta, J)$  estimators are substantially more efficient than the STP estimator: approximately 45% more efficient in estimating  $S(\tau|v0j0)$ , and 70% more efficient in estimating  $S(\tau|v1j1)$ . The greater magnitude of the gains for  $S(\tau|v1j1)$  reflects the fact that, in general, efficiency differences are due to the differential extraction of information from cases with unmeasured  $V_i$ . In this simulation approximately 8% of cases had V0J0, whereas 71% had V1J1. The differences in efficiency as a function of covariate values disappear when the simulations are set to produce equal number of cases in each covariate level (data not shown). Though for both covariate levels the RC-efficient are more efficient than the  $\hat{\pi}(\Delta, J)$ -estimators, these differences are small. Since for the saturated  $\hat{\pi}(\Delta, J)$ -estimator  $g_b = E \left[ D_i^{Fb} \mid J_i, \Delta_i \right]$  (Result 1, Appendix C), the slight advantage of the RC-efficient reflects the fact that little is gained by adding the actual observed time  $X_i$ , to the conditioning events. (INSERT TABLE 2 HERE)

Table 3 contains results for nonparametric estimators of  $S(\tau|v)$  when  $J_i$  is an auxiliary covariate rather than a risk factor. For example,  $J_i$  might be a surrogate for  $V_i$ , such as evidence of gastric inflammation found on a biopsies obtained at the beginning of the study. These simulations reveal two important features of estimation. First, they demonstrate the potential for gaining efficiency by utilizing auxiliary information: the  $\hat{\pi}(\Delta, J)$ -estimator (logistic model 17) is more efficient than the  $\hat{\pi}(\Delta)$ -estimator (logistic model 16). In simulations (not shown) where  $V$  and  $J$  are independent, the efficiency of the  $\hat{\pi}(\Delta, J)$ -estimator is identical to that of  $\hat{\pi}(\Delta)$ -estimator. Second, the contrast in the performance of the two different locally efficient estimating procedures illustrates some noteworthy properties of LE-estimators. Each of the corresponding named *simple (SLE)* and *insured (ILE) local efficient* estimators use identical estimates,  $\hat{g}_1^{2eff}$ , of  $g_1$ . However SLE-estimates are produced by setting  $g_1 = \hat{g}_1^{2eff}$  in (20), whereas ILE-estimates are  $\hat{\pi}$ -estimates based on prediction model (30) with  $g_1 = \hat{g}_1^{2eff}$ . By construction, ILE-estimators must be at least as efficient as  $\hat{\pi}(\Delta, J)$ -estimators, even when  $\hat{g}_1^{2eff}$  is based on a misspecified  $rr(X_i|v1)$ . SLE's do not share this property. For example, the  $\hat{g}_1^{2eff}$  of the *SLE and ILE correct-estimators* is based on a correctly specified models for  $rr(X_i|v1)$ . Specifically, we assumed exponential hazards within each  $V$  level; estimated the hazards by dividing the number of observed cases by total person-time; and estimated  $rr(X_i|v1)$  as a ratio of the hazards. Both the SLE and ILE correct estimators attain the nonparametric efficiency bound. In contrast the SLE and ILE *prior* and *null* estimators use misspecified  $rr(X_i|v1)$ 's. The *prior* estimators set  $rr(X_i|v1) = 0.5$ , the pooled estimate of  $rr$  from the prior studies. The *null* estimators set  $rr(X_i|v1) = 1$ ; these would be the efficient estimator under the null hypothesis. Table 3 shows that for estimators of  $S(\tau|v0)$  the SLE-prior estimator is less efficient than the  $\hat{\pi}(\Delta, J)$ -estimator. In simulations with  $V$  and  $J$  independent (data not shown), the SLE-prior has a variance 9% greater than even the STP-estimator. Thus the RE of the SLE estimators are not bounded above by 1. In

contrast, in the Table 3 simulations the ILE prior is more efficient than the  $\hat{\pi}(\Delta, J)$ -estimator. Under independence the efficiencies are nearly identical. Note that all locally efficient estimators, misspecified or not, are unbiased and have confidence intervals that cover at the stated rate. (INSERT TABLE 3 HERE)

## 10. DISCUSSION

In this paper we characterized the class of non and semiparametric cumulative hazard estimators, and, by extension, survival estimators, in two-stage studies. We formulated the difference between estimators within a class in terms of differences in their utilization of information from subjects on whom the stage-two exposures were not measured. In particular, we showed that the minimum variance estimators required estimating the expected value of the full data influence function conditional on the partially observed, e.g., stage-one, data. We expressed the requirements for estimating (or approximating) this conditional expectation in terms of concepts familiar to epidemiologists: efficient estimation requires knowledge about which covariates are independent risk factors for exposure and for disease. We discussed the impact of these considerations in structuring the design and analysis of two-stage studies. We emphasized a general approach to analysis,  $\hat{\pi}$ -estimation, which allows investigators to incorporate their knowledge in ways that can increase the efficiency of estimation without undermining consistency. Computer code in S-plus and R for these  $\hat{\pi}$ -estimators is given in Appendix F.

## References

- Andersen P.K., Borgan O., Gill R.D., and Keiding N. (1991), *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.
- Blaser M.J. (1999), "Hypothesis: The changing relationship of *Helicobacter pylori* and humans: implications for health and disease," *Journal of Infectious Diseases*, 179, 1523-1530.
- Blot WJ, Li JY, Taylor PR, Guo W, Dawsey S, Wang GQ, Yang CS, Zheng SF, Gail M, Li GY, Yu Y, Liu BQ, Tangera J, Sun YH, Lie FS, Fraumeni JF, Zhang YH, Li B. (1993), "Nutrition intervention trials in Linxian, China: supplementation with specific vitamin/mineral combinations, cancer incidence, and disease-specific mortality in the general population," *Journal of the National Cancer Institute*, 85,1483-1492
- Borgan O., Goldstein L., and Langholz B. (1995), "Methods for the analysis of sampled cohort data in the cox proportional hazards model," *The Annals of Statistics*, 23, 1749-1778.
- Dawsey S.M., Mark S.D., Taylor P.R., and Limburg P.J. (2002), "Gastric Cancer and H Pylori." *Gut*, 51, 457-458.
- Fleming, T.R., and Harrington D.P. (1991), *Counting Process and Survival Effects*, New York, Wiley.
- Helicobacter and Cancer Collaborative Group. (2001), "Gastric cancer and Helicobacter Pylori: a combined analysis of 12 case-control studies nested within prospective cohorts," *Gut*, 3, 347-353.
- Ihaka, R. and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299-314.
- Lidell F.D.K., McDonald J.C., and Thomas D.C. (1997), "Methods for cohort analysis: appraisal by application to asbestos mining (with discussion)," *Journal of the Royal Statistical Society Ser. A*, 140, 469-490.
- Limburg P.J., Wang C.Q., Mark S.D., Qiao Y.L., Perez-Perez G.I., Blaser M.J., Taylor P.R., Dong Z.W., and Dawsey S.M. (2001). "Helicobacter Pylori Seropositivity: Association with Increased Gastric Cardia and Non-Cardia Cancer Risks in Linxian, China," *Journal of the National Cancer Institute*, 93, 226-233.
- Mark S.D., Qiao Y.L., Dawsey S.M., Katki H., Gunter E.W., Yan-Ping W., Fraumeni J.F., Blot W.J., Dong Z.W., and Taylor P.R. (2000), "Higher serum selenium is associated with lower esophageal and gastric cardia cancer rates," *Journal of the National Cancer Institute*, 92, 1753-1763.

Mark S.D., and Katki H. (2001), "Influence function based variance estimation and missing data issues in case-cohort studies," *Lifetime Data Analysis*, 7, 329-342.

Newey W.K. (1990), "Semiparametric Efficiency Bounds", *Journal of Applied Econometrics*, 5, 99-135.

Prentice R.L. (1986), "A case-cohort design for epidemiologic cohort studies and disease prevention trials," *Biometrika*, 73,1-11.

Pugh M.G. (1993), "Inference in the Cox Proportional Hazards Model with Missing Covariate Data", unpublished Sc.D dissertation, Harvard School of Public Health, Boston, MA.

Robins J.M., Rotnitzky A., and Zhao L.P. (1994), "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89, 846-866.

Rubin D.B. (1976), "Inference and Missing Data," *Biometrika*, 63, 581-592.

Self S.G., and Prentice R.L. (1988), "Asymptotic distribution theory and efficiency results for case-cohort studies," *The Annals of Statistics*, 16, 64-81.

Therneau, T. M. and Grambsch, P. M. (2000), *Modeling Survival Data: Extending the Cox Model*, Springer-Verlag Inc.

### Appendix A.

In this appendix, we provide consistent estimators of the asymptotic variances for the two-stage estimators of cumulative hazards, relative risks, and survivals. When we can do so without confusion, and to indicate that any consistent estimator of a parameter will suffice, we drop the arguments  $g_b$ . For instance, we write  $\tilde{\Lambda}(\tau)$  for  $\tilde{\Lambda}(\tau, g_1)$ . We further define  $d\tilde{M}_i(u) = dN_i(u) - Y_i(u) d\tilde{\Lambda}(u)$ ;  $d\tilde{M}_i(u, \beta) = dN_i(u) - Y_i(u) d\tilde{\Lambda}_o(u, \tilde{\beta}) \exp(\tilde{\beta} Z_i)$ ;  $\tilde{s}^j(u, \cdot) = n^{-1} \tilde{S}^j(u, \cdot)$ ;  $\tilde{e}(u, \beta) = \tilde{s}^1(u, \beta) \tilde{s}^0(u, \beta_o)^{-1}$ ;  $\tilde{i} = n^{-1} \sum_{i=1}^n \pi_{i,o}^{-1} R_i \Delta_i \left( Z_i - \tilde{E}(X_i, \beta) \right) \left( Z_i - \tilde{E}(X_i, \beta) \right)^T$ ;

#### A.1. Estimating $D_i^b(g_b)$ (26), and $D_i^b(\hat{\pi}(W^l))$ (31)

Estimators  $\tilde{D}_i^b(g_b)$  of  $D_i^b(g_b)$  are formed by the obvious substitutions for  $s^j(u, \cdot)$ ,  $dM_i(u, \cdot)$ , and  $\tilde{e}(u, \beta)$  in 21, 23, 25. The weights  $\pi_{i,o}$  can be replaced by any consistent estimate,  $\hat{\pi}$ . For  $\hat{\pi}$ -estimators,  $\tilde{D}_i^1(\hat{\pi}(W^l))$  and  $\tilde{D}_i^2(\hat{\pi}(W^l))$  are formed by estimating  $P^{bl}$  (28) by the vector of regression parameters from an ordinary least squares regression of  $\tilde{D}_i^b(\pi_{i,o})$  on the scores  $\tilde{S}_i^l$ . Letting  $\hat{g}_{i,2}(\pi) = \pi_{i,o} P^2 W_i^l$ , we estimate  $D_i^3(\hat{\pi}(W^l))$  by substituting regression estimates from sequential least squares in the influence function.

$$D_i^3(\hat{\pi}(W^l)) \equiv D_i^3(g_2(\hat{\pi}), g_3^* = 0) - E[D_i^3(g_3^* = 0, g_2(\hat{\pi})) S^l] E[S_i^l S_i^{l'}]^{-1} S_i^l.$$

#### A.2 Estimating the asymptotic variance of $\tilde{\Lambda}(\tau)$ , and $\{\tilde{\beta}^T, \tilde{\Lambda}_o(\tau, \tilde{\beta})\}^T$

Let  $\tilde{D}_i^a = \{\tilde{D}_i^{2T}, \tilde{D}_i^3\}^T$ , and  $V_1$  and  $V_a$  be the variances of  $\tilde{\Lambda}(\tau)$  and  $\{\tilde{\beta}^T, \tilde{\Lambda}_o(\tau, \tilde{\beta})\}^T$  respectively. Consistent estimates of the asymptotic variance are  $\tilde{V}_1 = n^{-1} \sum \tilde{D}_i^1 \tilde{D}_i^{1T}$  and  $\tilde{V}_a = n^{-1} \sum \tilde{D}_i^a \tilde{D}_i^{aT}$ .

#### A.3 Estimating the asymptotic variance of $\tilde{S}(\tau|vj)$ .

Let  $\tilde{S}(\tau)$  and  $\tilde{S}(\tau, \beta)$  be the  $k^* \times 1$  vector of nonparametric and semiparametric estimates of  $S(\tau)$ , with row  $h$  entry  $\tilde{S}(\tau; h)$  and  $\tilde{S}(\tau; h, \beta)$ . Let  $V_{s1}$  and  $V_{s2}$  be the corresponding  $k^* \times k^*$  variance matrices for  $\tilde{S}(\tau)$  and  $\tilde{S}(\tau, \beta)$ . Define  $G$  as the  $k^* \times k^*$  diagonal matrix with  $\tilde{S}(h)$  in the  $h$ 'th row  $h$ 'th column. Then  $\tilde{V}_{s1} = G \tilde{V}_1 G_1$  is a consistent estimate of  $V_{s1}$ . Each  $h$  in the support of  $Z$  can be represented as a unique  $p \times 1$  covariate vector,  $Z_h$ . Let  $L_h = \tilde{S}(\tau; h, \beta) \exp(\tilde{\beta}^T Z_h) \times \{1, \tilde{\Lambda}_o(\tau, \beta) \times Z_h\}$ . Let  $L$  be the  $k^* \times (p+1)$  matrix with  $h$ 'th row  $L_h^T$ . Then  $\tilde{V}_{s2} = L \tilde{V}_a L^T$  is a consistent estimator of  $V_{s2}$ .

#### A.4 Estimating the asymptotic variances of $\tilde{S}(\tau|vj)$ , $\tilde{S}^s(\tau|v)$ , $\tilde{R}d(\tau)$ , and $P\tilde{A}R$ .

Consistent with common usage we define *standardized survival*,  $S^s(\tau|v)$ , to be the weighted sum of covariate specific survivals, with known weights,  $w(j^*)$ , which sum to 1. That is,  $S^s(\tau|v) = \sum_{\mathcal{J}} S(\tau|vj^*) w(j^*)$ .

Let  $v^*, j^*$  be the number of levels of  $V$  and  $J$  respectively. Arrange  $\tilde{S}(\tau|vj)$  in  $v^*$  groups of length  $j^*$ , in order of increasing index. Let  $W_j^T$  be the  $1 \times j^*$  matrix of weights  $w_j$ ;  $I_{v^*}$  the  $v^* \times v^*$  identity matrix; and  $C_w = W_j^T \otimes I_{v^*}$  where  $\otimes$  denotes the Kronecker product. Then  $\tilde{S}^s(\tau|v) = C_w \tilde{S}(\tau, \cdot)$  with variance estimated by, for instance,  $C_w \tilde{V}_{s2} C_w^T$ . Estimates of standardized risk differences,  $\tilde{R}d(\tau)$ , are simple contrasts of the  $\tilde{S}^s(\tau|v)$ . To estimate populztion attributable risks, PAR, we define the  $1 \times k^*$  "crude weight matrix",  $W_{vj}^T = \{w_{v1j1}, w_{v1j2}, \dots, w_{v1j^*}, w_{v2j1}, \dots, w_{v2j^*}, \dots, w_{v^*j1}, \dots, w_{v^*j^*}\}$ , with  $w_{vj} = \tilde{n}_{vj}/\tilde{n}$ ;  $\tilde{n}_{vj} = \sum_{i=1}^n R_i I(V_i = v, J_i = j)$   $\hat{\pi}_i^{-1}(W_i)$ ,  $\tilde{n} = \sum_{vj} \tilde{n}_{vj}$ . The estimator of the "crude population survival" is  $\tilde{S}^m(\tau) = W_{vj}^T \tilde{S}(v, j)$ . We let

$1_{vh} = \{0, 0, \dots, 1, \dots, 0, 0\}$  be the  $v^* \times 1$  vector with the only non-zero element being 1 in the  $h$ 'th row.. Then using  $v = v^\dagger$  as the unexposed  $V$  level we estimate the PAR as

$$P\tilde{A}R = \left( \tilde{S}^s(\tau|v^\dagger) - \tilde{S}^m(\tau) \right) \times \left( 1 - \tilde{S}^m(\tau) \right)^{-1} \\ = \{1_{v^\dagger}^T C_j - W_{vj}^T\} \vec{S}(\tau|v, j) \times \left( 1 - W_{vj}^T \vec{S}(\tau|v, j) \right)^{-1}$$

Using the delta method we obtain and estimate of the variance of the  $P\tilde{A}R$ ,  $\tilde{V}_{par} =$

$\partial \tilde{V}_{s2} \partial'$ , where  $\partial = \left( 1 - W_{vj}^T \vec{S}(\tau|v, j) \right)^{-1} \{C_{v^\dagger} - W_{vj}^T\} I_{k^*} + \{C_{v^\dagger} - W_{vj}^T\} \vec{S}(\tau|v, j) \left( 1 - W_{vj}^T \vec{S}(\tau|v, j) \right)^{-2} W_{vj}^T I_{k^*}$ , and  $I_{k^*} = k^* \times k^*$  identity matrix.

### Appendix B

In this appendix we prove that the variance of (24) is minimized when  $g_2 = 0$ ,  $g_3^* = E[D_i^{F3}|W_i]$ . Let  $C_{i1}(g_2) = \pi_{io}^{-1} R_i D_i^{F3} + i^{-1} \pi_{io}^{-1} (R_i - \pi_{io}) g_{i2} k_1$ ;  $k_1 = \int_0^\tau e(u, \beta_o) d\Lambda_o(u)$ . Then (25) is  $D_i^3(g_2, g_3^*) = C_{i1}(g_2) - \pi_{io}^{-1} (R_i - \pi_{io}) g_{i3}^*$ . For any fixed  $g_2$ , proposition 2.3 RRZ establishes that the  $g_3^*$  minimizing the variance of  $D_i^3(g_2, g_3^*)$  is,  $g_{i3}^{eff}(g_2) = E[C_{i1}(g_2)|W_i] = E[D_i^{F3}|W]$ . Hence the variance of  $D_i^3(g_2, g_3^*)$  is minimized by finding the  $g_2$  that minimizes

$$var C_{i,1}(g_2^*) - 2 cov[C_{i,1}(g_2^*), (\pi_{io}^{-1} (R_i - \pi_{io}) g_{i,3}^{eff})] \tag{B.1}$$

Taking the expectations in (B.1) conditional on  $W_i$ , the only term containing  $g_2$  is

$$E[\pi_{io}^{-1} (1 - \pi_{io}) (i^{-1} g_{i,2} k_1)^2]$$

which is minimized by  $g_{i,2} = 0$ .

### Appendix C

In this appendix we show that the variance of  $D_i^b(W^l)$  (27) is minimized iff  $q^b S_i^l = \pi_{io}^{-1} (R_i - \pi_{io}) E[D_i^{Fb}|W_i]$ . For any given set of scores,  $S_i^l$ , the variance is minimized when  $q^b = P^{bl}$  (28). Since the variance is non-increasing in the dimension of  $S_i^l$ ,  $P^{bl} S_i^l = \pi_{io}^{-1} (R_i - \pi_{io}) g_b^{eff}$  iff, for all  $W_i^m > W_i^l$ ,

$$E\left[ \left( D_i^b(\pi_{i,o}) - P^{bl} S^l \right)' S^{ml} \right] = 0. \tag{C.1}$$

Here  $S_i^{ml}$  are the linearly independent matrix of scores from the residual of the projection of  $S_i^m$  on  $S_i^l$ . Taking the expectation of (C.1) conditional on  $H_i$ , and using MAR restriction (14), this becomes  $E\left[ W_i^{ml} (1 - \pi_{io}) \right] E\left[ D_i^{Fb} - \pi_{io} P^{bl} W_i^l | W_i \right] = 0$ , which is true iff  $P^{bl} S_i^l = \pi_{io}^{-1} (R_i - \pi_{io}) E\left[ D_i^{Fb} | W_i \right]$ .

**Result 1:** Taking each of the expectations in  $P^{bl}$  conditional on  $H_i$ , we obtain  $P^{bl} = E$

$\left[ (1 - \pi_{io}) D_i^{Fb} \times (W_i^l)' \right] \times E\left[ \pi_{io} (1 - \pi_{io}) W_i^l W_i^l \right]^{-1}$ . Let  $W^f$  have discrete covariate space of dimension

$f^*$ , with model (15) parameterized so that the design matrix is the  $f^* \times f^*$  identity matrix. Then, the matrix of scores are orthonormal, and  $P^{bf} S_i^f = \pi_{io}^{-1} (R_i - \pi_{io}) E \left[ D_i^{fb} \middle| W_i^f \right]$ .

**Result 2:** To see that for (30)  $P^{bf} S_i^f = \pi_{io}^{-1} (R_i - \pi_{io}) g_b^{eff}$ , note that (30) is correctly specified with  $\psi_1 = 1, \psi_2 = 0$ ; by the general form of the  $P^{bl}$  in Result 1, the projection of  $D_i^b(\pi_{i,o})$  onto  $\pi_{io}^{-1}(R_i - \pi_{io}) W_{ib}^{eff}$  is  $\pi_{io}^{-1} (R_i - \pi_{io}) E \left[ D_i^{fb} \middle| W_i \right]$ . Since the span of the scores from model (30) is greater than the span of  $\pi_{io}^{-1} (R_i - \pi_{io}) W_i^{eff}$ , (29) is true for the scores from model (30).

### Appendix D

In this appendix we derive a general expression for  $K_{v^\dagger}(W_i)$ , and the specific expression given in (39).

We define  $\lambda_m^*(s|v^\dagger, A_i) = \lim_{h \rightarrow 0} Pr(s \leq X_i < s+h, \Delta_i = m | X_i \geq s, v^\dagger, A_i) / h$ ; (D.1)

$$\lambda_X^*(s|v^\dagger, A_i) = \sum_{m=0}^1 \lambda_m^*(s|v^\dagger, A_i) \tag{D.2}$$

$$rr_m(s|v^\dagger, A_i) = \lambda_m(s|v^\dagger, A_i) / \lambda_m(s|v^1, A_i) \tag{D.3}$$

where  $m \in \{0, 1\}$ . We further define a "non-informative censoring" assumption

$$Pr(C_i \geq s | T_i \geq s, V_i, A_i) = Pr(C_i \geq s | T_i \geq s, A_i) \tag{D.4}$$

The hazards in (D.1) are referred to as the crude hazards of  $C_i$ , ( $m = 0$ ), and  $T_i$ , ( $m = 1$ ). (D.2) is the hazard for the random variable  $X_i$ . Under the conditional independence assumption, the crude hazards equal the net hazards (e.g. Andersen et al.,

1991). Non-informative censoring assumption (D.4) is similar in subject matter content to the usual non-informative censoring assumption, but does not imply that assumption: the latter allows for dependency of both  $C_i$  and  $T_i$  on  $V_i$ , but requires that, in terms of factorability of the likelihood, such dependency be distinct (Andersen et al., 1991).

By Bayes' rule,

$Pr(v^\dagger | X_i, \Delta_i, A_i) = Pr(X_i, \Delta_i | v^\dagger, A_i) \times Pr(v^\dagger | A_i) / Pr(X_i, \Delta_i | A_i)$ . Writing  $Pr(X_i = x_i, \Delta_i | v, A_i)$  as  $\lambda_{\Delta_i}(x_i | v, A_i, X_i \geq x_i) \times Pr(X_i \geq x_i | v, A_i)$ , we obtain

$$K_{i,v^\dagger} = rr_{\Delta_i}(x_i | v^\dagger, A_i) \times Pr(X_i \geq x_i | v^\dagger, A_i) / Pr(X_i \geq x_i | v^1, A_i) \times Pr(v^\dagger | A_i) / Pr(v^1 | A_i) \tag{D.5}$$

Applying (D.4), the right hand side of (D.5) gives

$$rr_1(x_i | v^\dagger, A_i)^{\Delta_i} \times Pr(T_i \geq x_i | v^\dagger, A_i) / Pr(T_i \geq x_i | v^1, A_i) \times Pr(v^\dagger | A_i) / Pr(v^1 | A_i) \tag{D.6}$$

Using LE estimators as in section 9, one could postulate models for the distribution of  $T_i$  conditional on  $\{V_i, A_i\}$  and estimate (D.6); with additional assumptions about the conditional distribution of  $C_i$ , (D.5) can be similarly estimated. When  $J$  contains all the independent risk factors in  $A_i$ , (D.6) becomes (39).

Though (D.5-D.6, 38) are true regardless of the support of  $V_i$  and  $A_i$ , consistency of the estimators given in section 9 depends on the discreteness of  $A_i$ . When the support is not discrete,  $K_{i,v}$  can be approximated by forming discretized random variables  $A_i^s$ , and using the empirical distribution of  $Pr(V_i | A_i^s)$  instead of  $Pr(V_i | A_i)$ .

### Appendix E: $\hat{\pi}$ -estimators for Case-Cohort and Nested Case-Control Designs

In this appendix we provide  $\hat{\pi}$ -estimators when sampling follows that defined by either the CCH or NCC designs. For simplicity we assume sampling does not depend on  $A_i$ . Though both designs specify that  $V_i$  be observed on all cases, the  $\hat{\pi}$ -estimators we give require no such restriction. We assume only that cases are sampled with some known (dependent, or independent, probability). For detailed descriptions of sampling procedures see, for instance, Self and Prentice (1988), or Borgan et al. (1995).

In the CCH the "comparison" group is a binomial random sampling drawn from all cohort members. Since both the case and controls sampling probabilities are dependent only on  $\Delta_i$ , any  $\hat{\pi}$ -estimators with column space greater than (8) can be used.

NCC designs use dependent, risk set, sampling. Let  $\{\mathbf{T}_{(1)}, \dots, \mathbf{T}_{(d)}\}$  be the set of ordered case failure times. We estimate the case-sampling probability,  $\pi_{io}(\Delta_1)$ , by the proportion of cases sampled. For subjects

with  $\Delta_i = 0$ , we define indicator variables,  $R_{ik} = 1$ , if the subject is selected at  $T_{(k)}$ ; and  $\bar{R}_{ik} = 1$ , if  $R_{ih} = 1$ , for some  $h \leq k$ ;  $\bar{R}_{i0} \equiv 0$ . Let  $\pi_{i,k} \equiv Pr(R_{ik} = 1 | X_i, \Delta_i = 0, \bar{R}_{i,k-1} = 0)$ , then

$$Pr(R_i = 1 | \Delta_i = 0, X_i) \equiv \pi_{io}(\Delta_0) = \sum_{k=1}^d \pi_{ik} I(X_i \geq T_{(k)}, \bar{R}_{i,k-1} = 0) \prod_{j=1}^{k-1} (1 - \pi_{ij}) \quad (E1)$$

where the product term is defined to be 1 when  $k = 1$ . To estimate  $\pi_{io}(\Delta_0)$ , we replace the  $\pi_{i,k}$  in (E1) with the proportion of controls with  $(X_i \geq T_{(k)}, \bar{R}_{i,k-1} = 0)$  who were sampled at  $T_{(k)}$ . Estimating the influence function requires obtaining scores from the likelihood based on  $\pi_{io}(\Delta)$ .

**Appendix F: Computer Code for Implementing the  $\hat{\pi}$ -estimators in R 1.70 or S-plus 6.0 Release 2**  
( appendix F was written by Steven D. Mark and Hormuzd Katki)

To facilitate correspondence with the data analysis and simulations the code given is written for the structure of the data described in Sections 7 and 8, and implements the non and semiparametric  $\hat{\pi}(\Delta, J)$ -estimators (17). Making the appropriate changes to the design matrix in the code in F1 below, allows implementation of  $\hat{\pi}$ -estimators for any correctly specified logistic model (15,18). The semiparametric example we give has no auxiliary covariates, so that  $J_i = A_i$ . The general structure of the code poses no such constraint. That is, the covariates used to predict  $\hat{\pi}$ , need not be the same as those specified by the CPH model in F3. In the code below we refer to the set of ordered case-failure times  $\{T_{(1)}, \dots, T_{(d)}\}$ , and to  $\tilde{N} \cdot(u) = \sum \pi_{io}^{-1} R_i N_i(u)$ .

$\tilde{N} \cdot(u)$  estimates the full data aggregated counting process,  $N \cdot(u) = \sum \pi_{io}^{-1} R_i N_i(u)$ . The semiparametric estimator assumes the hazards are specified by CPH model (4).

**F1. Obtain predicted selection probabilities,  $\hat{\pi}(W_i)$  (15-17) and scores,  $S_i$  (19) from logistic regression models.**

**F1.1** Attach the MASS library. This allows us to use the generalized inverse to estimate the  $E[S S']^{-1}$  required for the projection operator,  $P^{bl}$ . The generalized inverse is necessary if, for example, there is some  $W_i$  such that  $\pi(W_i) = 1$ . This would occur if all cases were sampled.

```
library(MASS) )
```

**F1.2** Fit the  $\hat{\pi}(\Delta, J)$  selection model (17). The covariate matrix "x" in the code, corresponds to the  $W^l$  in (17). We keep the covariate matrix by setting the option x=True.

```
options(contrasts=c("contr.treatment","contr.poly"))
propmod <- glm(R ~ J*Delta, x=True, epsilon=1e-10, family=binomial,
data=hpdata)
```

**F1.3** Obtain the estimated  $\hat{\pi}(\Delta, J)$ : pihat.

```
pihat = predict(propmod,type='response')
```

**F1.4** Obtain the scores,  $S_i^l$ , (19) from the prediction model.

```
scores <- matrix(R-pihat,nrow=n,ncol=length(propmod$coeff)) * propmod$x
```

**F2. Nonparametric estimators of survival  $S(\tau|vj)$**

In this section we provide code for estimating  $\tilde{S}(\tau|vj)$ , and the variance of  $\tilde{S}(\tau|vj)$  for the nonparametric  $\hat{\pi}(\Delta, J)$ -estimator. F2.1-2.2 returns survivals for all  $k^* = 4$  strata at each observed failure time in the stratum. F4.3 retrieves the survival  $\tilde{S}(\tau|v0j0)$ , for "stratum 1", the stratum with  $V = 0, J = 0$ . Here  $\tau$  is the time of the last observed failure time,  $T_d$ , in stratum 1. F2.5-2.8 estimate the influence function (21, A.1) for the  $\hat{\pi}(\Delta, J)$  estimator of the cumulative hazard for stratum 1. The influence function is called D1.phat.1. F2.9 provides the variance estimator for  $\tilde{S}(\tau|v0j0)$ . To obtain the variance matrix for all 4 survival estimators

requires cycling through #F2.3-2.8 and calculating the stratum specific survivals,  $\text{surv.k}$ , and influence functions,  $D1\text{phat.k}$ ,  $k \in \{1, 2, 3, 4\}$ . F2.10 produces the covariance matrix  $\tilde{V}_1$  of  $\tilde{\Lambda}(\tau)$ , and  $\tilde{V}_{s1}$  of  $\tilde{S}(\tau)$  as described in A.2 and A.3.

Note that though we obtain the survival estimate by using the S-PLUS 6.0 (R 1.6.2) function `survfit()`, the variances returned by the `survfit()` function are not consistent for either the STP or  $\hat{\pi}$ -estimators. In our experience, the variance estimates from `survfit()` are considerably smaller than the true variances.

**F2.1** Place `pihat` in the dataset; select subjects in `stratum1: V0,J0`.

```
hpdata$pihat <- pihat
stratum1 <- hpdata[R==1 & V==0 & J==0,]
```

**F2.2** Estimate  $S(u|v, j)$  with weights  $\hat{\pi}(\Delta, J)^{-1}$  at each of the  $d$  event times in each of the  $k^*=4$  stratum.

```
hpsurv <- survfit(Surv(X,Delta==1)~J+V, type="fl", weights=1/pihat,
na.action=na.omit,data=hpdata)
```

**F2.3** Obtain the estimate  $\tilde{S}(\tau|v_0j_0)$ . The variable, `risksetofinterest`, is the index of the last event time in stratum 1.

```
risksetofinterest <- length(hpsurv[1]$n.event) -
which(rev(hpsurv[1]$n.event)>0)[1] + 1
print(surv.1 <- hpsurv[1]$surv[risksetofinterest])
```

**F2.4** Obtain  $\tilde{S}^0(u)$  (Section 4) and the aggregated counting process,  $\tilde{N}^0(u)$ , for stratum 1.

```
time=0,S0(time=0),dNdot(time=0)=0.
failures <- hpsurv[1]$n.event!=0
time <- c(0,hpsurv[1]$time[failures])
S0 <- c(sum(1/stratum1$pihat),hpsurv[1]$n.risk[failures])
dNdot <- c(0,hpsurv[1]$n.event[failures])
```

**F2.5** Compute `riskset`, the index of the last failure time at risk for each subject in stratum 1.

```
riskset <- matrix(0,nrow=dim(stratum1)[1],ncol=1)
for (i in 1:dim(stratum1)[1]) {
eligible <- rev(time <= stratum1$X[i])
riskset[i] <- (length(time)+1)-which(eligible==max(eligible))[1]}
```

**F2.6** Estimate  $D_{i,1}^1(\pi_{io})$ , the influence function for the STP estimator of  $\tilde{\Lambda}(\tau)$  (21, A.1), for stratum 1. Note  $\tilde{D}_{i,1}^1(\pi_{io}) \equiv 0$  for individuals not in stratum 1 (e.g. `stratum1=0`).

```
correction <- cumsum(S0^-2 * dNdot)
influence <-
(1/stratum1$pihat)*((1/S0[riskset])*(stratum1$Delta==1)*
(riskset<=risksetofinterest) - correction[pmin(riskset,risksetofinterest)])
D1 <- matrix(0,nrow=n,ncol=1)
D1[R==1 & V==0 & J==0] <- influence
```

**F2.7** Estimate the  $\hat{\pi}(\Delta, J)$  offset,  $P^1 S_i$  (29, A.1): `gamma1`.

```
gamma1 <- scores %*% ginv(t(scores)%*%scores) %*% (t(scores) %*% D1)
```

**F2.8** Estimate the variance of  $\tilde{\Lambda}(\tau)$  for stratum 1 (A.2).

```
D1pihat.1 <- D1-gamma1
print(varcumhaz <- t(D1pihat.1)%*%(D1pihat.1))
```

**F2.9** Estimate the variance of  $\tilde{S}(\tau|vj)$  (A.3) for stratum 1.

```
print(versurv <- surv^2 * varcumhaz)
```

**F2.10** To obtain the overall variance matrix (A.3) of the survivals in all  $k^* = 4$  strata, repeat #4.3-4.8 as described at the start of section 4.

```
G <- diag(c(surv.1, surv.2, surv.3, surv.4))
D1pihat <- matrix(c(D1pihat.1, D1pihat.2, D1pihat.3, D1pihat.4),ncol=4)
print(V.s1 <- t(G) %*% t(D1pihat) %*% D1pihat %*% G)
```

### F3 Estimating $\beta_o$ in CPH model (4)

The code produces consistent estimates for the  $\hat{\pi}$ -estimators of  $\beta_o$ , and the variance of the estimators. The predicted probabilities, pihat, are obtained from the output of F1. We estimate  $P^2S_i$  (29, A.1) by a code that is analogous, though not identical, to that provided for S-PLUS in Chapter 7.3 of Therneau and Grambsch (2000). The main difference of importance is that we use the generalized inverse for reasons described in Section 3.

**F3.1** Estimate  $\beta_o$  with weights  $\hat{\pi}(\Delta, J)^{-1}$ . (NOTE: robust=TRUE or cluster() do not yield the correct variance)

```
coxmod <-
coxph(Surv(X,Delta==1)~J+V,weights=1/pihat,method="breslow",
na.action=na.omit,x=T,data=hpdata)
print(beta.til<- coxmod$coef)
```

**F3.2** Estimate  $D_i^2(\pi_{io})$  (D2 below), the STP influence function (A.2).

Note,  $\tilde{D}_i^2(\pi_{io})$  is a  $p \times 1$  vector. When  $R_i = 0$ ;  $\tilde{D}_i^2(\pi_{io}) \equiv 0$ .

```
D2 <- matrix(0,nrow=n,ncol=length(coxmod$coeff))
D2[R==1,] <- residuals(coxmod,'dfbeta',weighted=T)
```

**F3.3** Estimate the  $\hat{\pi}(\Delta, J)$  offset,  $P^2S_i$ : gamma2.

```
gamma2 <- scores %*% ginv(t(scores)%*%scores) %*% (t(scores) %*% D2)
```

**F3.4** Estimate the variance of  $\tilde{\beta}$ .

```
D2pihat <- D2-gamma2
print(varbeta <- t(D2pihat)%*%(D2pihat))
```

### F4 Semiparametric estimator of survival $\mathcal{S}(\tau|vj)$

In this section we provide code for the  $\hat{\pi}(\Delta, J)$ -estimator of the survivals,  $\tilde{\mathcal{S}}(\tau, \beta)$  and their variances,  $\tilde{V}_{s2}$  (A.3). The  $\tilde{\mathcal{S}}(\tau, \beta)$  is a vector of dimension  $k^* = 4$ . The beta.til, and D2 are obtained from the output of F3. The influence function of the  $\hat{\pi}$ -estimator of  $\Lambda_o(\tau, \beta)$  is calculated by first obtaining the influence function for the STP-estimator (called D3 in the code) in F4.4. In F4.5 we subtract off the correct "offset" and obtain the influence function of the  $\hat{\pi}(\Delta, J)$ -estimator (called D3pihat).

**F4.1** Add to the data set the estimates of  $\pi_{i,o}$  and  $\beta_o$  from section 3 and 5. Select only individuals with  $R_i = 1$ .

```
beta.til <- coxmod$coef
hpdata$pihat <- pihat
fulldata <- hpdata[R==1,]
```

**F4.2** Obtain  $\tilde{S}^0(u, \beta)$ , and the estimator of the aggregated counting process  $N(u)$ . Note that the hazard of coxph.details, is the "centered hazard",  $\lambda_o \exp \tilde{\beta}^T \bar{Z}(u)$ , where  $\bar{Z}(u)$  is the  $p \times 1$  vector of covariate averages of the individuals at risk at time  $u$ .

```
coxdetails <- coxph.detail(coxmod)
time <- c(0,coxdetails$time)
hazard <- coxdetails$hazard/exp(apply(coxdetails$x,2,mean) %*% beta_til)
Ndot <- cumsum(coxdetails$weights[coxdetails$y[,3]==1])
  [!rev(duplicated(rev(coxdetails$y[coxdetails$y[,3]==1,2])))]
dNdot <- c(0,Ndot[1],diff(Ndot))
S0 <- c( sum(coxdetails$weight * exp(coxdetails$x %*% beta_til) ), dNdot[-1] /
hazard )
```

**F4.3** Compute riskset, the index of the last failure time at risk for each subject.

```
riskset <- matrix(0,nrow=dim(fulldata)[1],ncol=1)
for (i in 1:dim(fulldata)[1]) {
eligible <- rev(time <= fulldata$X[i])
riskset[i] <- (length(time)+1)-which(eligible==max(eligible))[1]}
```

**F4.4** Estimate,  $D_i^3(\pi_o)$ , the influence function for the STP-estimator. Here D1 is an estimator of the first term,  $\pi_{io}^{-1} R_i \int_0^{\tau} [s^0(u, \beta_o)]^{-1} dM_i(u)$ , of that influence function. For individuals with  $R_i = 0$ , D1=0.

```
risksetofinterest <- length(S0)
correction <- cumsum(S0^2 * dNbar)
D1 <- matrix(0,nrow=n,ncol=1)
D1[hpdata$R==1,] <-
(1/fulldata$pihat)*((1/S0[riskset])*fulldata$Delta==1)*(riskset<=risksetofinterest)
- correction[pmin(riskset,risksetofinterest)]*exp(coxmod$x %*% beta)
)
E <- coxdetails$mean
EdLambda <- apply( E * matrix(hazard,nrow=length(hazard),ncol=2),2,sum )
D3 <- D1 - ( D2 %*% EdLambda )
```

**F4.5** Estimate  $D_i^3(\hat{\pi})$  (A.1), the influence function for the  $\hat{\pi}(\Delta, J)$ - estimator of  $\Lambda_o(\tau, \beta_o)$ .

```
gamma1 <- scores %*% ginv(t(scores)%*%scores) %*% (t(scores) %*% D1)
D3pihat <- (D3 -gamma1
print(varBaselineCumhaz <- t(D3pihat) %*% D3pihat)
```

**F4.6** Estimate  $\tilde{V}_{s2}$  (A.3) the  $k^* = 4$  vector of the variance of the survival estimates,  $\tilde{S}(\tau, \beta)$ . Here Z are the 4 unique covariate levels,  $Z = [0\ 0], [0\ 1], [1\ 0], [1\ 1]$ .

```
Z <- matrix(c(0,0,0,1,1,0,1,1),nrow=4,ncol=2,byrow=T)
print(surv.beta<-exp(-cumhaz*exp(Z %*% beta)))
D.a <- matrix(c(D3pihat,D2pihat),ncol=3,byrow=F)
V.a <- t(D.a) %*% D.a
L <- matrix(surv.beta*exp(Z %*% beta),nrow=4,ncol=3) *
matrix(c(1,1,1,1,cumhaz*Z),nrow=4,ncol=3)
print( V.s2 <- L %*% V.a %*% t(L) )
```

**Table 1.**  
**Estimating Survival Conditional on Hp and Age at 5.25 years in the Linxian Cohort**

|            | Survival (95% CI)       |                         |
|------------|-------------------------|-------------------------|
|            | <i>H. Pylori</i> - (V0) | <i>H. Pylori</i> + (V1) |
| Young (J0) | 99.2 (98.9, 99.5)       | 98.8 (98.4, 99.0)       |
| Old (J1)   | 97.3 (96.1, 98.1)       | 95.5 (94.4, 96.3)       |

The estimates are based on the CPH model with relative risk  $\exp(\beta_1 V + \beta_2 J)$ .  
 The  $\hat{\pi}(\Delta, J)$ -estimator (17) was used for estimating  $\beta_o$  and  $\Lambda_o(\tau, \beta_o)$ .

**Table 2.**  
**The STP, RC-efficient, and  $\hat{\pi}(\Delta, J)$  Semiparametric Estimators of  $S(\tau | v, j)$**

| Estimator              | $S(\tau   v0, j0) = 90\%$ |                 |                     | $S(\tau   v1, j1) = 73.5\%$ |                 |                     |
|------------------------|---------------------------|-----------------|---------------------|-----------------------------|-----------------|---------------------|
|                        | Mean Survival<br>V=0, J=0 | 95% CI Coverage | Relative Efficiency | Mean Survival<br>V=1, J=1   | 95% CI Coverage | Relative Efficiency |
|                        | STP                       | 95.0            | 94.7                | 100                         | 73.5            | 95.5                |
| RC-efficient           | 95.0                      | 95.0            | 55                  | 73.6                        | 95.7            | 27                  |
| $\hat{\pi}(\Delta, J)$ | 95.0                      | 95.6            | 57                  | 73.5                        | 95.2            | 31                  |

Note: Relative efficiency equals 100 times the ratio of the variance of the estimator to the variance of the STP estimator.

**Table 3.**  
**The STP,  $\hat{\pi}(\Delta, J)$ , SLE, and ILE non-parametric estimators of  $S(\tau | v)$  in the presence of an auxiliary covariate**

| Estimator              | $S(\tau   v0) = 90\%$ |                 |                     | $S(\tau   v1) = 81\%$ |                 |                     |
|------------------------|-----------------------|-----------------|---------------------|-----------------------|-----------------|---------------------|
|                        | Mean Survival<br>V=0  | 95% CI Coverage | Relative Efficiency | Mean Survival<br>V=1  | 95% CI Coverage | Relative Efficiency |
| STP                    | 90.0                  | 94.1            | 100                 | 81.0                  | 94.7            | 100                 |
| $\hat{\pi}(\Delta)$    | 90.0                  | 94.5            | 83                  | 81.0                  | 95.4            | 50                  |
| $\hat{\pi}(\Delta, J)$ | 90.0                  | 93.8            | 74                  | 81.0                  | 94.8            | 45                  |
| SLE correct            | 90.0                  | 94.4            | 71                  | 81.0                  | 95.7            | 40                  |
| ILE correct            | 90.0                  | 94.4            | 71                  | 81.0                  | 95.7            | 40                  |
| SLE prior              | 90.0                  | 94.9            | 85                  | 81.0                  | 95.4            | 43                  |
| ILE prior              | 90.0                  | 94.2            | 71                  | 81.0                  | 95.5            | 40                  |
| SLE null               | 90.0                  | 94.4            | 74                  | 81.0                  | 95.7            | 40                  |
| ILE null               | 90.0                  | 94.4            | 71                  | 81.0                  | 95.8            | 40                  |

Note: Relative efficiency equals 100 times the ratio of the variance of the estimator to the variance of the STP estimator