

Submitted Journal American Statistical Association, November 2003

Specifying and Implementing Nonparametric and Semiparametric Survival Estimators in Two-Stage (sampled) Cohort Studies with Missing Case Data

Steven D. Mark and Hormuzd A. Katki

Steven D. Mark is a Senior Research Investigator, and Hormuzd A. Katki is a Staff Scientist, in the Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd., EPS Room 8036, Bethesda, MD, 20852-7368. *e-mail*: smark@exchange.nih.gov. The authors thank Jamie Robins for many helpful discussions.

Abstract

Since 1986 we have been studying a cohort of individuals from a region in China with epidemic rates of gastric cardia cancer. To assess the association of various exposures with this cancer we have conducted numerous two-stage studies. Two-stage studies are a commonly used statistical design: stage-one consists of observing the outcomes and accessible baseline covariate information on all cohort members; stage-two consists of using the stage-one observations to select a subset of the cohort for measurements of exposures that are difficult to obtain. When the outcomes are censored failure times such as in our studies, the most common designs used are the case-cohort and nested case-control designs (Samet and Munoz 1998). One limitation of both these designs is that the estimators of the cumulative hazards, and hence survival and absolute risk, are biased when some cases are missing the second stage measurements (Mark and Katki 2001; Mark 2003). In our cohort the exposures of interest are potentially responsive to population-wide interventions. Hence survival estimates are crucial to public health decisions. In all of our studies some cases are missing measurements, either by chance or by design. In this paper we present a class of nonparametric and semiparametric cumulative hazard estimators that are unbiased regardless of case sampling fraction. We analyze data from a study of the association of gastric cardia cancer with serologic evidence of *H. pylori* infection in which we sampled only twenty-five percent of available cases. We estimate differences in cancer incidence between individuals with and without infection. Simulations based on this study demonstrate that a wide variation in efficiency exists between estimators within a class. We characterize the mathematical form of the efficient estimators within each class; show the practical implications for study design and analysis; and provide strategies for choosing an efficient estimator. Computer code in R or S-plus for implementing these estimators is available from the authors.

KEY WORDS: absolute risk; auxiliary covariate; case-cohort; cumulative hazard; efficiency; nested-case control; risk difference; robust estimation; survival; standardized survival; two-stage studies; weighted estimating equations.

Introduction

Nearly all large epidemiologic cohort studies initiated in the last thirty years have been designed to estimate the association of a disease with exposures measured on biologic samples (Samet and Munoz 1998). Since measurements on such samples are typically expensive and consume scarce resources, a class of statistical designs has risen where the goal is to reduce the number of measurements while still obtaining estimates with adequate precision. These designs, commonly called two-stage studies by statisticians (Robins, Rotnitzky, and Zhao 1994; heretofore RRZ), and "nested cohort studies" by epidemiologists (Samet and Munoz 1998), have the following general structure. At the start of the cohort, time 0, investigators obtain biological specimens as well as measurements on a large number of other covariates from each cohort member. Endpoints of interest are recorded up until some time, τ . We refer to this collected covariate and endpoint data as the *stage-one data*, and designate it as W_i . In stage two, the W_i observed on all n subjects in the cohort are used to select a subset of individuals on whom measurements on biological samples will be made. We call those measurements, V_i . For example, in the data analysis we present in section 6, $V_i = 1$ if an individual has serological evidence of *H. pylori* (Hp) infection; $V_i = 0$ otherwise. Understandably all two-stage designs call for sampling a smaller fraction of controls (individuals without the observed endpoint) than cases. In fact, since statistical efficiency is largely determined by the number of observed cases, it is generally specified that two-stage designs select for V_i measurement "all the cases of interest, but only a subsample of the noncases" (Samet and Munoz 1998, p 8). Choosing a particular two-stage design requires consideration of the endpoint of interest, e.g. cumulative disease incidence, or survival time; the risk parameter of interest, e.g., odds ratios, relative risks, or absolute risks (survival probabilities); and the underlying models. In epidemiology the latter generally entail specifying a regression model relating outcomes to the exposures, V_i , and to adjusting covariates, J_i . For cumulative incidence endpoints this is commonly a logistic model; for censored failure time data a Cox Proportional hazards model (CPH) such as in (1). RRZ give a comprehensive review of the different statistical features of the proposed two-stage designs. Samet and Munoz (1998) discuss the implementation and motivation for these designs in specific settings, and present some of the important findings produced by these studies.

We focus on studies where the endpoint is censored failure time. Specifically, in section 6 we analyze data in which the event of interest is the time, T_i , to the development of gastric cardia cancer (GCC). Rather than T_i , we observed the right censored event outcome (X_i, Δ_i) , where $X_i = \min(T_i, C_i)$, C_i is an independent censoring time, $\Delta_i = I(X_i = T_i)$, and $I(\cdot)$ is the indicator function. As in most large cohorts, nearly all censoring was due to censoring by the end of follow-up at time, $\tau = 5.25$ years. We refer to individuals with $\Delta_i = 1$ as cases; those with $\Delta_i = 0$ as controls. Though a number of two-stage designs have been proposed for estimation with censored failure time endpoints, nearly all are variations of the original nested-case control (NCC) (Borgan, Goldstein, and Langholz 1995) and case-cohort (CCH) proposals (Prentice 1986; Self and Prentice 1988). The NCC and CCH designs are by far the most common designs used in practice (Samet and Munoz 1998). Both assume the CPH model; both specify that V_i be measured on all cases. The main distinction between the NCC and CCH designs is the control sampling schemes; we briefly review these in Appendix B. For a comprehensive discussion of these and other related sampling schemes for two-stage studies with censored time-to-event outcomes see Mark and Katki (2001).

The primary focus of the NCC and CCH designs has been estimation of the $p \times 1$ vector of parameters, $\beta_o = \{\beta_{o1}^T, \beta_{o2}^T\}^T$, in a CPH model such as (1)

$$\lambda(u|Z_i) = \lambda_o(u) \exp(\beta_{o1}^T V_i + \beta_{o2}^T J_i). \quad (1)$$

Here $Z_i = \{V_i, J_i\}$ is a p -dimensional vector of *exposure covariates*, V_i , and *adjusting covariates*, J_i , $J_i \subset W_i$. Frequently J_i contains information from questionnaires, physical exams, and/or measurements from laboratory procedures used to establish cohort eligibility, or to serve as baseline measures of attributes of interest. Though the emphasis has been on estimating β_o , both the NCC and CCH designs provide estimators of the cumulative hazards, $\Lambda(t; z)$

$$\Lambda(t; z) = \int_0^t \lambda(u|z) du \quad 0 \leq t \leq \tau; z \in \mathcal{Z} \quad (2)$$

where \mathcal{Z} is the support of Z_i . Just as estimates of relative risks ($rr(z)$) are obtained from the identity $rr(z) = \exp(\beta_o^T z)$, estimates of survival are obtained from the identity

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

One limitation of the proposed cumulative hazard estimators is that unlike the estimators of β_o , they are biased if any cases are missing V_i measurements (Mark and Katki 2001; Mark 2003). This

limitation was the motivation for the research we present. We have conducted a large number of two-stage studies on our cohort from Linxian, China (see section 6); in each some cases were missing V_i measurements. The etiology of the missing case measurements can be divided into two broad categories: missingness that occurs by *chance*; and missingness that occurs by *design* (Mark and Katki 2001; Mark 2003). Cases missing by chance arise from events outside of investigators' control. Indeed, due to the *missing by chance* mechanism, we suspect there are no large cohorts studies using biological specimens that can measure V_i on all cases. In contrast, we define cases to be *missing by design* if investigators deliberately measure V_i on only a fraction of all available cases. In section 6 we describe the reasons why estimating survival is crucial to the context of our research; the sources and magnitude of the chance missingness in our studies; and the limitations in biological resources which necessitated two-stage studies with cases missing by design.

Obtaining consistent survival estimators in two-stage studies with missing case information was the initial impetus for this research. As we will demonstrate, one can produce unbiased estimators of survival by simple inverse probability weighting of the subjects with measured V_i (Mark 2003). However, as verified by our simulations in section 7, substantial differences in efficiency exist between survival estimators all of which are unbiased. In particular, we will show that the simplest and best known inverse-probability-weighted estimator, the Horvitz-Thompson estimator (Horvitz and Thompson 1952), is so inefficient, and so easily improved upon, that we conclude that it should never be used. Hence in this paper we focus on the factors that determine the efficiency of estimators all of which are unbiased.

The organization of the paper is as follows. In section 2 we formally state the goals of our inference, and the structure of the two-stage studies we consider. In section 3 we define the term *full data estimators*; then, applying, the general results of RRZ on how to obtain two-stage estimators from full data estimators, we derive a class of unbiased semiparametric and nonparametric cumulative hazard estimators. For the former we assume that hazards are given by a CPH model (1); for the later we make no assumptions about hazards at different levels of covariates. In section 4 we define a general method for implementing our estimators, which we call $\hat{\pi}$ -estimation, and describe how the efficiency of one $\hat{\pi}$ -estimator relates to another. In section 5 we apply the theorems of RRZ to

censored time-to-event data and derive the mathematical form of the most efficient estimators within each class of estimators we consider. We re-express the efficient form in terms of quantities already familiar to researchers involved in observational studies, and show the general implications for study design and analysis. In section 6 we review features of several of our two-stage studies, and use a $\hat{\pi}$ -estimating procedure to estimate the effect of Hp infection on absolute risks and risk differences. In section 7 we apply the general formulation of what constitutes efficient estimators presented in section 5, and propose specific estimators. In simulations we demonstrate that the relative efficiency of these estimators correspond to predictions from theory. In the discussion section we provide a simple and non-technical summary of the results and their practical consequences. Annotated code in R (Ihaka and Gentleman 1996) and S-plus (6.0 release 2) that implements the $\hat{\pi}$ -estimating procedures is available from the authors (Mark and Katki 2003).

In this paper we restrict the mathematical results presented in Mark (2003) to that subset necessary for understanding the practical implications of the origins of the efficiency differences in unbiased estimators. Additionally, to make the presentation more accessible, in the body of the paper we express results only in terms of those functionals of the random variables required to understand the proofs and their importance to applications. Actually defining these functionals requires counting process and martingale notation. With the exception of two expressions in section 3.1 (these are not essential for any subsequent results), this notation is confined to the appendices. At points in the paper where some readers may desire more details or clarifications, we explicitly reference the appropriate sections in Mark (2003). There the derivation and discussion of these and other results are presented in a more general, more detailed, and more technical context.

2. Formal Statement of Inference, Data Structure, and Sampling Process

2.1 Standardized survival, standardized risk difference, full data, and auxiliary covariates

Our main goal of inference is to estimate conditional survivals (3), *standardized survivals*, $S^s(t|v^\dagger)$, and standardized risk differences. In accord with usual epidemiologic parlance, we define $S^s(t|v^\dagger)$ as the weighted sum over j of the $S(t|z)$, $z \in \mathcal{Z}$

$$S^s(t|v^\dagger) = \sum_j S(t|v^\dagger, j^*) w(j^*); \quad 0 \leq t \leq \tau. \quad (4)$$

Here v^\dagger and j^* represent specific points in the support of V_i and J_i ; the weights, $w(j^*)$, are functions of j^* chosen by the investigator which sum to 1. In the analysis of the Hp data the only adjusting covariate, J_i , is age, and we define $w(j^*)$ to be the observed marginal distribution of J_i in the cohort. Had we wished to compare disease incidence in Linxian with those in different geographic areas we could have used standardization weights for US, European, or world populations available at <http://seer.cancer.gov/stdpopulations/>. For a given set of weights, the standardized risk differences are a simple contrast

$$Rd(t) = S^s(t|v0) - S^s(t|v1) \quad (5)$$

where, for instance, we use $v0$ to represent $V_i = 0$. Since we wish primarily to make inference about survivals in groups of individuals, we assume that the support \mathcal{Z} is finite with k^* levels. Though cumulative hazards and survivals can be estimated at any time t , for simplicity we assume we are interested in these quantities at the end of the study, and for the remainder of the paper set $t = \tau$.

Were resource limitation not a factor, one could measure V_i on everyone and obtain "full data", $H_i = \{W_i, V_i\}$. To distinguish endpoint from covariate data, we write, $W_i = \{X_i, \Delta_i, A_i\}$. We assume the covariates in A_i are measured at time 0; A_i generally contains orders of magnitude more measurements than the set of adjusting covariates, J_i . Though while designing a cohort study it is essential to consider which adjusting covariates will be required, typically J_i is not formally specified until the analysis stage. Then J_i is usually chosen to be the subset of covariates in A_i that are known, or suspected, of being associated with T_i , and/or V_i , and are not "on the causal pathway". We refer to the (possibly empty) set of covariates that are in A_i but not J_i , as *auxiliary covariates*. Thus $A_i = \{J_i, \Lambda_i^{aux}\}$. The term auxiliary indicates that we do not wish to make inference about the cumulative hazards $\Lambda(t; z)$, conditional on Λ_i^{aux} . In a sense made more precise in section 5 and in the simulations, the Λ_i^{aux} can substantially increase the efficiency of estimation when they are correlated with V_i .

2.2. Stage-Two Sampling Restrictions

We define $R_i = 1$ if V_i is known for individual i ; $R_i = 0$ otherwise. 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

$$\pi_o(W_i) = Pr(R_i = 1 | W_i, V_i) = Pr(R_i = 1 | W_i). \quad (6)$$

In the usual parlance of missing data, restriction (6) is consistent with V_i being missing at random (MAR) (Rubin 1976). As we frequently do for random variables, we drop the 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 4 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 by

$$\text{logit } \pi_o(W_i) = \psi_o' h(W_i). \quad (7)$$

Here ψ_o and $h(W_i)$ are known, conformable, finite dimensional vectors of parameters and random variables, respectively. It is important to note that the function $h(\cdot)$ is not uniquely determined by the $\pi_o(W_i)$; that is, neither the parameterization nor the dimension of equation (7) 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 (7) are

$$\text{logit } \pi_o(W_i) = \psi_{o1} I(\Delta_i = 1) + \psi_{o2} I(\Delta_i = 0) \quad (8)$$

$$\begin{aligned} \text{logit } \pi_o(W_i) = \psi_{o1} I(\Delta_i = 1) + \psi_{o2} I(\Delta_i = 0) + \\ \psi_{o3} I(\text{male}) + \psi_{o4} I(\text{female}) \end{aligned} \quad (9)$$

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

The usefulness of models such as (9) will become evident when we discuss $\hat{\pi}$ -estimation in section 4.

We define W_i^R to be the smallest set of linearly independent vectors such that (7) is true, where size refers to the dimension of the column space spanned by $h(W_i)$ ($\text{span } h(W_i)$). In our example above, the dimension of W_i^R is two. Letting $W_i^l \equiv h(W_i)$ for some $h(\cdot)$, then correctly specified models are those such that

$$\text{span } (W_i^l) \supseteq \text{span } (W_i^R). \quad (10)$$

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 ,

$$S_i^l = (R_i - \pi_{i,o}) W_i^l. \quad (11)$$

As usual the S_i^l are the partial derivative with respect to ψ of the log likelihood of R_i .

3. Estimators and Influence Functions

3.1 Full data estimators and full data influence functions

Though our inferential focus is survival, cumulative hazards are the compensator of the counting process, and, therefore, the "natural" scale for estimation. For this and other details on counting process martingales, we refer the reader to Andersen, Borgan Gill, and Keiding, 1991.

In full data studies the Nelson-Aalen estimator, $\hat{\Lambda}(\tau; z)$, is the efficient nonparametric estimator of (2) (Anderson et al. 1991). The maximum partial likelihood estimator, $\hat{\beta}$, and the Breslow estimator, $\hat{\Lambda}_o(\tau, \hat{\beta})$ are the semiparametric efficient estimators of β_o (2) and the baseline cumulative hazard (12) (Anderson et al. 1991),

$$\Lambda_o(\tau) = \int_0^\tau \lambda_o(u) du . \quad (12)$$

For the semiparametric model we write the cumulative hazard at any covariate level z , by $\Lambda(\tau; \beta_o z) = \Lambda_o(\tau) \exp(\beta_o^T z)$. It is estimated in the obvious fashion. To indicate the $k^* \times 1$ vector of cumulative hazards, we drop z from the arguments and write $\Lambda(\tau)$, or $\Lambda(\tau; \beta_o)$. Since we are assuming the time of interest is τ , we frequently drop the time argument.

Letting $\hat{\alpha}^1$, $\hat{\alpha}^2$, denote the Nelson-Aalen and $\hat{\beta}$ estimators, we can, in a general sense, write these estimators as the $\hat{\alpha}^b$, $b \in \{1, 2\}$, that solve estimating equations of the form

$$\sum_{i=1}^n U_i^b(H_i, \mathcal{R}(X_i); \alpha^b) = 0 . \quad (13)$$

Each term in (13) depends not only on the subjects data, H_i , but also on $\mathcal{R}(X_i)$. $\mathcal{R}(X_i)$ represents the set of individuals at risk at time X_i ; e.g. $\mathcal{R}(X_i) = \{i : X_j \geq X_i\}$. For instance, using standard counting process notation (see A.1) when $b = 2$, then $\alpha^b = \beta$; $U_i^2(H_i, \mathcal{R}(X_i); \alpha^2) = \int_0^\tau \left\{ (Z_i) - S^1(u, \beta) S^0(u, \beta)^{-1} \right\} dN_i(u)$; and the maximum partial likelihood estimator, $\hat{\beta}$, is the β that solves $\sum_{i=1}^n \int_0^\tau \left\{ Z_i - S^1(u, \beta) S^0(u, \beta)^{-1} \right\} dN_i(u) = 0$.

For the Breslow estimator one first estimates $\hat{\beta}$. Then with $\hat{\alpha}^3$ denoting $\hat{\Lambda}_o(\tau, \hat{\beta})$, we can similarly write the estimating equations as

$$\sum_{i=1}^n U_i^3(H_i, \mathcal{R}(X_i); \hat{\beta}; \alpha^3) = 0 . \quad (14)$$

Note that the U_i^b are column vectors with row dimension k^* , p , and 1, for $b=1,2,3$ respectively. See section 2 of Mark (2003) for the explicit forms for $b = 1$ and $b = 3$.

Though the $U_i(\cdot)$ are not iid, the estimators are asymptotically equivalent to a sum of mean 0, independent, influence functions (Anderson et al. 1991). That is,

$$n^{\frac{1}{2}} \left(\hat{\alpha}^b - \alpha_o^b \right) = n^{-\frac{1}{2}} \sum_{i=1}^n D_i^{Fb}(H_i; \alpha_o^b) + o_p(1) \quad (15)$$

where α_o^b is the underlying parameter being estimated. We refer to these D_i^{Fb} , $b \in \{1, 2, 3\}$ as the *full data influence functions* of $\hat{\Lambda}(\tau)$, $\hat{\beta}$, and $\hat{\Lambda}_o(\tau, \hat{\beta})$, respectively. The D_i^{Fb} are functions of the observed data and the hazards; the explicit definitions are given in appendix A2 . For instance, for $b = 2$, A2 gives $D_i^{F2} = i^{-1} \int_0^\tau \{ Z_i - e(u, \beta_o) \} dM_i(u)$, and (15) becomes

$$n^{\frac{1}{2}} \left(\hat{\beta} - \beta_o \right) = n^{-\frac{1}{2}} \sum_{i=1}^n i^{-1} \int_0^\tau \{ Z_i - e(u, \beta_o) \} dM_i(u) + o_p(1).$$

3.2 Estimators and influence functions for two-stage designs

For two-stage designs, RRZ establish that the solutions to estimating equations

$$\sum_{i=1}^n \pi_{i,o}^{-1} R_i U_i^b(H_i, \mathcal{R}(X_i); \alpha^b) - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_b(W_i) = 0 ; b \in \{1, 2\} \quad (16)$$

produce consistent, asymptotically normal nonparametric and semiparametric estimators of the cumulative hazard (2) and β_o , respectively. Here the $g_b(W_i)$ are any conformable vector of non-stochastic functions of W_i specified by the investigator. We denote estimators based on a given g_b as $\tilde{\Lambda}(g_1); \tilde{\beta}(g_2)$. Similarly, solutions to

$$\sum_{i=1}^n \pi_{i,o}^{-1} R_i U_i^3(H_i, \mathcal{R}(X_i); \tilde{\beta}(g_2); \alpha^3) - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_3^*(W_i) = 0 \quad (17)$$

define a class of two-stage estimators of (12). We denote those estimators by $\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*)$, or $\tilde{\Lambda}_o(\tau, \tilde{\beta}, g_3)$. Here g_3^* is any scalar function of W_i , and $g_3(W_i)$ is the function of g_2 and g_3^* defined in (A4.1). The explicit estimating equations for (16) and (17) are given in (A3.1-A3.3) and in Mark (2003, section 4).

We write the influence functions that correspond to these classes of two-stage estimators as

$$D_i^b(g_b) = \pi_{i,o}^{-1} R_i D_i^{Fb} - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_b(W_i) , b \in \{1, 2, 3\}. \quad (18)$$

For $b \in \{1, 2\}$ (18) follows directly from RRZ. For $b = 3$, (18) is obtained by a Taylor series expansion around β_o (appendix A.4). Using notation analogous to the full data case (15), we express the two-stage estimators as a sum of their influence functions,

$$n^{\frac{1}{2}} \left(\tilde{\alpha}^b(g_b) - \alpha_o \right) = n^{-\frac{1}{2}} \sum_{i=1}^n D_i^b(g_b) + o_p(1) . \quad (19)$$

From (19) it is clear that the asymptotic variances of the $\tilde{\alpha}^b(g_b)$ are $E[D_i^b(g_b) D_i^b(g_b)^T]$.

Let $\tilde{\Lambda}(\tau, z, \cdot)$ be any nonparametric, e.g. $\tilde{\Lambda}(g_1; z)$, or semiparametric, e.g., $\exp(\tilde{\beta}^T(g_2)z) \times \tilde{\Lambda}_o(\tau, \tilde{\beta}, g_3)$, two-stage estimator of (2). Then survival estimates, $\tilde{S}(\tau|z)$, are formed by replacing $\Lambda(\tau; z)$ in (3) with $\tilde{\Lambda}(\tau, z, \cdot)$. Asymptotic distributions are derived by a straightforward application of the functional delta method exactly as in Andersen et al. 1991. We provide consistent estimators of the variances of (3), (4), and (5) in Appendix A.

4. The Simple True- π (STP) and $\hat{\pi}$ -Estimators

We define *simple true- π* (STP) estimators to be estimators where $g_{ib} \equiv 0$. That is, these are the usual, inverse-probability-weighted, Horvitz-Thompson estimators. However, rather than using the notation in (18) and writing these STP estimators as $D_i^b(g_b = 0)$, we denote their influence function by $D_i^b(\pi_o)$, which by (18) is

$$D_i^b(\pi_o) = \pi_{i,o}^{-1} R_i D_i^{Fb}$$

We define *$\hat{\pi}$ -estimating procedures* to be procedures in which we continue to set $g_{ib} \equiv 0$, but replace the known $\pi_{i,o}$ in estimating equations (16) and (17) with an estimate, $\hat{\pi}(W_i^l)$, of $\pi_{i,o}$. The predicted sampling probabilities, $\hat{\pi}(W_i^l)$, are obtained by replacing ψ_o with its maximum likelihood estimate, $\hat{\psi}$, in a correctly specified model (7) with covariates $h(W_i) = W_i^l$. We refer to estimators from such procedures as *$\hat{\pi}$ -estimators*. RRZ (proposition 6.2) showed that $\hat{\pi}$ -estimators are consistent, asymptotically normal, and that the influence function, $D_i^b(\hat{\pi}(W^l))$, is the residual of a population least squares regression of (20) on the scores from the prediction model, S_i^l . That is

$$D_i^b(\hat{\pi}(W^l)) = D_i^b(\pi_o) - P^{bl} S_i^l \quad (21)$$

and $P^{bl} S_i^l$, is the projection operator

$$P^{bl} = E[D_i^b(\pi_o) S_i^{l'}] E[S_i^l S_i^{l'}]^{-1}. \quad (22)$$

Since $D_i^b(\hat{\pi}(W^l))$ is a residual, the variance of the $\hat{\pi}(W^l)$ -estimator is less than or equal to the variance of the STP estimator for all W^l . Additionally, since residuals are non-increasing in the dimension of the column space, if $\text{span}(W_i^m) \supset \text{span}(W_i^l)$, the variance of the $\hat{\pi}(W^m)$ -estimator is less than or equal to the variance of the $\hat{\pi}(W^l)$ -estimator. For a more in-depth discussion of the properties of $\hat{\pi}$ -estimating procedures, and proof that the $\hat{\pi}$ -estimating procedures and the solutions to

estimating equations (16) and (17) generate the identical class of estimators, see Mark (2003, sections 5&6; appendix C).

$\hat{\pi}$ -estimation 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 and 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 $\text{span}(W_i^l) \supseteq \text{span} W_i^f$ (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 B. We have so far assumed that $\pi_{i,o}$, or equivalently, the ψ_o in logistic models (7), are known. If rather than knowing ψ_o , we only know there is a ψ^* , such that $\text{logit } \pi_{i,o} = \psi^* W_i^l$, then the estimator $\hat{\pi}(W^l)$ has influence function given by (21) (RRZ, proposition 6.2). For instance, to obtain consistent $\hat{\pi}$ -estimators for our Linxian studies we had to assume that we could correctly specify a logistic model that accounted for the chance missingness. Given the nature of the events causing the missingness (see section 6), we believed that missingness was related to neither W_i or V_i . Hence, any $\hat{\pi}$ -estimator with $\text{span}(W_i^l) \supseteq \text{span}(W_i^R)$ would be consistent.

Computer code for implementing the general class of $\hat{\pi}$ -estimating procedures is available from the authors (Mark and Katki 2003). This program handles a completely general data structure, and gives estimates, and the variances, for conditional survivals (3), standardized survivals (4), risk differences (5), and population attributable risks (for population attributable risk estimators and their asymptotic distribution see appendix A.4 of Mark 2003). We have used this program to produce nonparametric survival curves for a paper analyzing the association of zinc levels in biopsy tissue with esophageal cancer (Abnet et al. in press), and to produce semiparametric survival curves and risk estimates for the nutrient analyses described in section 6.

Section 5: Efficiency, Identifiability, and Local efficiency

5.1 Efficiency and the optimal g_b

Referring to $\pi_{i,o}^{-1}R_i$ as the *weight*, and $\pi_{i,o}^{-1}(R_i - \pi_{i,o})g_b(W_i)$ as the *offset*, it is clear from estimating equations (16) and (17), and influence functions (18), that the class of two-stage estimators we consider are "*weighted versions with offset*" of the efficient full data estimators. Since specific estimators differ only with regard to g_b , efficiency differences are determined entirely by the choice of the g_b function. We use g_b^{eff} to denote the optimal g_b : that is, the g_b that minimizes $E[D_i^b(g_b) D_i^b(g_b)^T]$. By results of RRZ (1994), and Newey (1990), who showed that all regular nonparametric full data estimators are asymptotically equivalent, the class of nonparametric estimators, $\tilde{\Lambda}(g_1)$ defined by (16, A3.1) contains (in the sense of asymptotic equivalence) all possible nonparametric cumulative hazard estimators for two-stage designs. Hence the estimator $\tilde{\Lambda}(g_1^{eff})$ achieves the nonparametric efficiency bound (Mark 2003). In contrast, for semiparametric estimators, we have followed a "practical recommendation" of RRZ (p850), and, restricted consideration to a subclass of all possible two-stage semiparametric estimators that use the "full-data efficient $h(\cdot)$ function" (Mark 2003). We refer the reader to RRZ for the general definition of the $h(\cdot)$ functions, and its specific form in two-stage estimators of (1). Thus we call estimators using the g_b that minimizes the variance of D_i^2 and D_i^3 , the restricted-class efficient (RC-efficient) estimators (Mark 2003).

For $b \in \{1, 2\}$, direct application of proposition 2.3 of RRZ establishes that $g_{i,b}^{eff} = E[D_i^{Fb} | W_i]$. By applying the same result to estimators of $\Lambda_o(s, \beta)$ with fixed g_2 , we find that $g_3^* = E[D_i^{F3} | W_i]$. It is simple to then show that the variance of $\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), E[D_i^{F3} | W_i])$ is minimized with $g_2 = 0$ (Mark 2003, appendix B). By definition of g_3 (A4.1), $g_3^{eff} = E[D_i^{F3} | W_i]$.

Replacing g_b with $\left[D_i^{Fb} \middle| W_i \right]$ in (18), demonstrates that these efficiency results correspond to intuition: every subject contributes $E\left[D_i^{Fb} \middle| W_i \right]$ to estimation; subjects with measured V_i provide the additional information in their observed "residual", $\pi_{i,o}^{-1}(D_i^{Fb} - E\left[D_i^{Fb} \middle| W_i \right])$.

In general we can express the influence function of the $\hat{\pi}$ -estimators (21) in terms of (18) by setting $g_{ib} = \pi_{i,o} P^{bl} W_i^l$. Using the double expectation rule, conditioning on H_i , and applying sampling restriction (6), we find that $P^{bl} = E\left[(1 - \pi_{i,o}) D_i^{Fb} \times (W_i^l)' \right] \times E\left[\pi_{i,o} (1 - \pi_{i,o}) W_i^l W_i^{l'} \right]^{-1}$. It is evident that for any $h(W_i) \supseteq W_i^R$, $\hat{\pi}$ -estimators based on predicted probabilities from logistic model (23)

$$\text{logit } \pi_o(W_i) = \psi_1' h(W_i) + \psi_2' W_i^{eff}; \quad W_i^{eff} = \pi_{i,o}^{-1} g_{i,b}^{eff} \quad (23)$$

are efficient, or RC-efficient (Mark 2003, appendix C).

5.2 Identification of g_b^{eff} , and implications for study design and analysis

The optimal $g(\cdot)$, $E[D_i^{Fb}|W_i]$, is a function of unknown parameters. RRZ proposition (2.4) established that g_b^{eff} can be replaced by a consistent estimate, \widehat{g}_b^{eff} , without changing the asymptotic distributions of two-stage estimators. That is, an estimator using \widehat{g}_b^{eff} achieves the efficiency, or RC-efficiency, bound. If g_b^{eff} can be consistently estimated, we say the efficient estimator is identified. If not, then variance of the efficient influence function represents an unknown lower bound that no estimator is guaranteed to achieve.

Were the support of W_i discrete, g_b^{eff} could be consistently estimated by the empirical average of the D_i^{Fb} among individuals with $R_i = 1$ within each level of W_i ; a $\widehat{\pi}$ -estimator saturated in the discrete W_i obtains that efficiency bound. In time to event data, W_i has the continuous component, X_i . Unless X_i were a deterministic function of (Δ_i, A_i) , there is no discrete subset $W_i^l \subset W_i$ such that $E[D_i^{Fb}|W_i^l] = E[D_i^{Fb}|W_i]$. In the remainder of this section we approach the task of conditioning on X_i and increasing efficiency. We do this 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 $g_{i,b}^{eff}$ (Mark 2003) as

$$g_{i,b}^{eff} = EE[D_i^{Fb}|W_i, V_i] = \int_{\mathcal{V}} D_i^{Fb}(W_i, v) Pr(v|W_i) dv. \quad (24)$$

In the design stage, a crucial consideration is what, if any, auxiliary variables should be measured.

From (24) it is clear that for Λ_i^{aux} to be optimal, it is sufficient that for any larger set, $\Lambda_i^{aux+} > \Lambda_i^{aux}$,

$$Pr(v|X_i, \Delta_i, J_i, \Lambda_i^{aux}) = Pr(v|X_i, \Delta_i, J_i, \Lambda_i^{aux+}). \quad (25)$$

That is, we should collect all auxiliary information that provides additional knowledge about the distribution of the incompletely measured covariates V_i at any time on study. Letting v^1 be some reference level of interest in \mathcal{V} , we can parameterize the $Pr(v|W_i)$ in (24) in terms of the "exposure odds"

$$K_{i,v} = Pr(V_i = v|W_i) / Pr(V_i = v^1|W_i). \quad (26)$$

Using Bayes' rule, $Pr(v|W_i) = Pr(X_i, \Delta_i|v, A_i) \times Pr(v|A_i) / Pr(X_i, \Delta_i|A_i)$, and a non-informative censoring assumption, $Pr(C_i \geq s|T_i \geq s, V_i, A_i) = Pr(C_i \geq s|T_i \geq s, A_i)$, (26) becomes (Mark 2003, appendix D)

$$K_{i,v} = rr(X_i|v, A_i)^{\Delta_i} \left(S(X_i|v, A_i) / S(X_i|v^1, A_i) \right) \left(Pr(v|A_i) / Pr(v^1|A_i) \right) \quad (27)$$

Here $rr(X_i|v, A_i)$ and $S(X_i|v, A_i)$ are conditional relative risks and survival times when the conditioning event is (V_i, A_i) rather than the (V_i, J_i) . Then (25) is true if

$$S(u|V_i, J_i, \Lambda_i^{aux}) = S(u|V_i, J_i, \Lambda_i^{aux+}); 0 \leq u \leq \tau \quad (28)$$

and

$$P(V_i|J_i, \Lambda_i^{aux}) = P(V_i|J_i, \Lambda_i^{aux+}). \quad (29)$$

Epidemiologists refer to (28) as Λ_i^{aux} containing all *independent predictors of outcome*; and (29) as Λ_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 Λ_i^{aux} that contains the independent predictors of outcome and exposure.

5.3 Efficient and locally efficient estimators

Though for any given Λ_i^{aux} it is impossible to know with certainty whether (28) or (29) is true, these are the exact considerations required to control confounding. As described in section 2, J_i is frequently defined in the analysis stage to be the subset of A_i such that (28) and (29) are "approximately" true when Λ_i^{aux} is removed from the conditioning events on the left hand side. If successful in selecting all of the disease risk factors in J_i then

$$S(u|V_i, J_i) = S(u|V_i, J_i, \Lambda_i^{aux}), 0 \leq u \leq \tau \quad (30)$$

and (27) becomes

$$K_{i,v} = rr(X_i|v, J_i)^{\Delta_i} \times S(X_i|v, J_i) / S(X_i|v^1, J_i) \times Pr(v|A_i) / Pr(v^1|A_i). \quad (31)$$

The identifiability and efficiency results we give in this section assume (30) is true.

From (31) it is clear that if we can consistently estimate each term in $K_{i,v}$ we can estimate \hat{g}_b^{eff} . For both the nonparametric and semiparametric models, the second and third terms can be estimated by $\tilde{S}(X_i|v^\dagger, J_i)$ and $\hat{P}(v^\dagger|A_i)$, the empirical average of V_i within levels of A_i (here we assume A_i has finite support; see Appendix D Mark 2003 for the case where the support of A_i is not finite). For the

semiparametric model, $rr(u|Z_i)$ can be estimated by $r\tilde{r}(u|Z_i) = \exp(\tilde{\beta}^T Z_i)$. The $\tilde{S}(X_i|v^\dagger, J_i)$ and $\tilde{\beta}$ can come from estimates based on any g_2, g_3^* . Hence, the semiparametric RC-efficient estimators of β_o and $\Lambda_o(\tau)\exp(\beta_o^T Z_i)$ 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 7.3 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^{LE} . If the assumptions about the rr 's are correct, then \hat{g}_b^{LE} is a consistent estimate of g_b^{eff} , and the LE-estimators are efficient. Regardless of the truth of the assumptions, the proposed *LE-estimators* are consistent.

Section 6: Two-stage studies conducted on the Linxian Cohort: Goals, Constraints, and Data Analysis

6.1 Two-stage studies with cases missing by chance

Since 1986 we have been studying a cohort of approximately 30,000 individuals from Linxian, China, a region with epidemic rates of GCC cancer (Blot and Li 1985; Blot et al. 1993). The cohort was assembled to investigate the hypothesis that one or more of the widely prevalent nutrient deficiencies contributed to this high GCC incidence. After following the cohort for 5.25 years and recording data on incident GCC and censoring events, we initiated four major studies where the V_i were measurement(s) of a group of related nutrients (Mark et al. 2000; Mark et al. 2001; Abnet et al. 2003, Taylor et al. 2003). We wanted to estimate nutrient-GCC associations with as much precision as possible, so for these studies our design called for sampling one-hundred percent of the 402 incident GCC cases. Despite the fact that virtually one-hundred percent of our cohort consented to giving blood at the beginning of the study in 1986, we discovered that accidents in sample processing, storage, shipping, or laboratory evaluation, prevented measurements for approximately 10% of the cancers (Mark et al. 2000). Using the standard case-cohort estimators of relative risk we found that serum levels of selenium and vitamin E were inversely related to cancer incidence (Mark et al. 2001; Taylor et al. 2003). The strongest effect was for selenium where individuals in the highest quartile of selenium had approximately half the cancer risk of those in the lowest (Mark et al. 2001). A variety of strategies for population wide nutrient supplementation to eliminate these deficiencies are currently

being considered by our colleagues at the Cancer Institute, Chinese Academy of Medical Sciences. Decisions of whether and how much to supplement depends on estimates of absolute risks. Using a $\hat{\pi}$ -estimating procedure we estimated that the correction of both selenium and vitamin E deficiencies could reduce the GCC incidence by approximately 30%. We are currently preparing a manuscript describing these results.

Many of the two-stage studies we have initiated in the last four years have examined the association of GCC with recently characterized DNA polymorphisms (Stolzenberg-Solomon et al. 2003; Savage et al. in press; Roth et al. in press; Mahabir et al. submitted). Samples suitable for DNA measurements were not collected until 1991, and then only on a subgroup of the remaining cohort. Overall we measured polymorphisms in approximately 20% of the cases from 1991-1996. Thus, in these studies 80% of the cases were missing by chance.

6.2 "Exploratory" two-stage studies where cases are missing by design

By the time we designed the serological studies of nutrients, numerous other exposures that could be measured in serum had become of interest. Since our total serum quantity was quite limited, and the list of exposures of interest large, we initiated "exploratory" two-stage studies in which we deliberately sampled only a fraction of cases (Abnet et al. 2001; Limburg et al. 2001). Our goal was to sample only the number of cases and controls required to produce precise enough estimates of exposure prevalence, assay reliability, and risk magnitude to decide whether to commit additional resources (Mark and Katki 2001). In one "preliminary study" where the exposure was the fungal-produced toxin fumonosin, we found the newly developed measurement procedure was not reliable, and have not initiated a larger study (Abnet et al. 2001). In contrast, due to the results from the "exploratory" study on the association of GCC with serologic evidence of *H. pylori*, we have begun a much larger two-stage study.

6.3 Background information on the association of Hp with GCC

Cancers that arise in the proximal 2-3 centimeters of the stomach are called gastric cardia cancers (GCC). These differ with regard to population rates, and some individual level risk factors, from stomach cancers that arise outside of the cardia (GNC) (Devesa, Blot, Fraumeni 1998). In the last decade epidemiologic cohort studies have found that individuals with Hp infection are at increased

risk of GNC; relative risks (rr 's) range from two to four (Helicobacter and Cancer Collaborative Group 2001). The quantity, consistency, and biologic plausibility of the evidence is such that Hp is categorized as a class 1 human carcinogen (International Agency for Research on Cancer 1994).

Prior to our study, only a few small studies, case sizes ranging from 4-12, examined the Hp-GCC association. All were from first-world Western nations. The consensus was that Hp was "protective" for GCC, with $rr \approx 0.5$ (Helicobacter and Cancer Collaborative Group 2001; Dawsey, Mark, Taylor, et al. 2002). Various mechanistic hypothesis have been advanced to account for the opposite association of Hp on GNC and GCC (Blaser 1999).

6.4 Design and analysis of the Hp-GCC study using a cohort from Linxian, China

Based on dissimilarities between the populations, and on differences in the prevalence of esophageal adenocarcinomas which can reduce the accuracy in diagnosing GCC (Limburg et al 2001; Dawsey et al. 2002), we hypothesized that the Hp-GCC association in Linxian might differ from that found in Western populations. In accord with the goals for "exploratory" studies given above, we sampled approximately 25% of the GCC cases (100 cases) and 7% of controls (200 controls) that occurred in the cohort by $\tau = 5.25$ years (Limburg et al. 2001). We measured serum antibodies and found an Hp prevalence, (Hp^+ , $V_i = 1$) 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, ($J_i = 1$) increased GCC risk by a factor of 3.5.

Table 1 contains estimates of covariate specific survivals (3), age standardized survivals (4), and risk differences (6) based on the CPH model (1) with V_i and J_i indicator variables. Since a fixed number of cases ($n=200$) and controls ($n=100$) were sampled, we used a $\hat{\pi}$ -estimator to estimate both β_o and $\Lambda_o(\tau)$. In particular, we used logistic model (9), the model saturated in (Δ_i, J_i) . Throughout this paper we denote this estimator by $\hat{\pi}(\Delta, J)$. 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 (J_0). We estimated the standardized risk difference to be 1.08, with a 95% confidence interval whose lower limit just excludes zero.

We contributed the data from our study to a pooled study examining Hp and gastric cancer risks. The overall conclusion of that analysis was that there was no evidence of an Hp-GCC

association (Helicobacter and Cancer Collaborative Group 2001). We did not share that interpretation. Rather we argued that tests for heterogeneity of risk estimates by geographic region were highly significant (Dawsey et al. 2002), and that pooling the risk estimates from Western populations and Chinese populations was not appropriate. We have currently initiated a larger study sampling from the approximately 1000 incident GCC that have accrued through 2001 ($\tau = 15$ years). This is also a study where cases are missing by design; however here the motivation for the designed missingness is opposite to that described above. Based on the Hp prevalence and risk estimates from the "exploratory" study, we determined that measurements on all 1000 GCC were not needed to achieve the precision required to eliminate type 1 error as a viable explanation for our earlier findings. The simulations in section 7 are based on the structure of this new study. Similar simulations were used to help arrive at the sampling fractions used in the actual study.

Section 7: Simulations

7.1 Simulation parameters and definition of relative efficiency

For all simulations the marginal covariate probabilities were $Pr(J1) = 0.5$, and $Pr(V1) = 0.65$; T_i was specified by CPH model (1); $\lambda_o(u)$ was exponential; censoring was independent. The magnitudes for the baseline hazard and competing risks were chosen to produce approximately 1000 expected cases in a cohort of size $n=6600$ by time τ . Unless otherwise noted, $\exp(\beta_{o1}) = 2$ ($rr_v = 2$). The V-J association was altered by changing the conditional probabilities, $P(V1|J1)$. Stage 2 sampling was binomial, and depended only on case status (8). Control sampling was 15%. For the simulations in Tables 2 and 3, 25% of cases were sampled, resulting in a control to case ratio of approximately 3:1. In Figures 1 and 2 case-sampling percents are indicated along the x -axis.

Each of the results represents the average of 2000 realizations. All estimators of survivals and β_o were unbiased: the mean of the estimators was always within 0.1% of the truth. The coverage for 95% confidence intervals ranged from 93.4% to 95.8%. Consequently, rather than present the estimator specific averages in the tables, we report only relative efficiencies (RE), which we define 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. Since our focus is on survival estimation, we do not report the RE's of estimators of β_o .

7.2 STP, RC-efficient, and $\hat{\pi}$ semiparametric estimators of survival

The data in Table 2 were generated from CPH model (1) with $\beta_{o2} = 0$; $S(\tau|v)$ was estimated by fitting the one-covariate CPH model, $\lambda_o(u) \exp(\beta_{o1}V_i)$. For simulations on the left hand side of Table 2, $rr_v = 2$; $S(\tau|v0) = 90$; $S(\tau|v1) = 81$. For simulations on the right, $rr_v = 0.5$, $S(\tau|v0) = 90$; $S(\tau|v1) = 95$. In these simulations J_i is an auxiliary variable 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 biopsy obtained at the beginning of the study. We compare $\hat{\pi}(\Delta)$ -estimators based on logistic model (8), with the $\hat{\pi}(\Delta, J)$ -estimator based on (9) at five different levels of V-J association. We focus first on the $rr_v = 2$ simulations.

When $Pr(V1|J1) = 0.65$, V_i and J_i are independent. Hence the $\hat{\pi}(\Delta)$ and $\hat{\pi}(\Delta, J)$ estimators are equally efficient, and considerably more efficient than the STP estimator. Since the $\hat{\pi}(\Delta)$ estimator makes no use of the auxiliary variable, its RE is unchanged as $Pr(V1|J1)$ increases. In contrast, the efficiency of the $\hat{\pi}(\Delta, J)$ estimator increases (RE decreases).

Differences in the efficiencies between two-stage estimators is largely determined by the extent to which information from cases with $R_i = 0$ is used. In the $rr_v = 2$ simulations, the magnitude of the efficiency gains for both $\hat{\pi}$ -estimators is greater in the $v1$ than $v0$ strata. These greater gains reflect the fact that there are more cases, and hence more cases with missing measurement, in the $v1$ stratum. When $rr_v = 2$, 78% of the missing cases occur in the $v1$ group. In contrast, when $rr_v = 0.5$, 49% of the missing cases occur in the $v1$ stratum. Here both strata have nearly identical RE's. Contrasting the two sets of simulations with regard to estimation of $S(\tau|v1)$, we find lower RE's when $rr_v = 2$. The expected number of missing cases for the $rr_v = 2$ simulation is 775; for the $rr_v = 0.5$, the expected number is 209.

The comparable nonparametric $\hat{\pi}$ -estimators for the Table 2 simulations produced the same patterns and are not shown. Variances of the nonparametric estimators were five to ten percent larger than their semiparametric counterparts.

The first two rows of Table 3 contain results for the RC-efficient estimator, which we denote by \tilde{S}^{eff} , and the $\hat{\pi}(\Delta, J)$ -estimator of standardized survivals for the semiparametric model (1) with $\beta_{o1} = 2$, $\beta_{o2} = 3$. As expected, the RE's of both estimators are less than one, with the RE of

\tilde{S}^{eff} lower than that of $\hat{\pi}(\Delta, J)$. In Figure 1 we present the relative efficiencies of the \tilde{S}^{eff} and $\hat{\pi}(\Delta, J)$ estimators of $S^s(\tau|v1)$ at four different case sampling probabilities: 12.5%, 25%, 50%, or 90%. The positive slope of the line indicates the efficiency gains decrease as sampling fraction increases. At 12.5% case sampling the RE of the \tilde{S}^{eff} estimator is 35%; at 90% case sampling the RE is 59%. Similarly, differences in efficiency between the \tilde{S}^{eff} (solid line) and $\hat{\pi}(\Delta, J)$ (dotted line) estimators diminish with increasing case-sampling percentage. Both these findings accord with the prior observation that efficiency gains depend on the number of missing cases. The same explanation accounts for the greater efficiency gains for estimators of $S^s(\tau|v1)$ compared to estimators of $S^s(\tau|v0)$ in Table 3. Eighty-five percent of the expected 1023 cases have $V_i = 1$.

In Figure 2 the solid line is a plot of the ratio of the variance of the \tilde{S}^{eff} estimator of $S^s(\tau|v1)$ at each of the four case-sampling probabilities, to the variance of the STP estimator at 90% sampling. The RC-efficient estimator has a lower variance when 25% of the cases are sampled, than the STP estimator has when 90% of the cases are sampled. The dotted line compares the variance of \tilde{S}^{eff} at each case-sampling percent, to the variance of \tilde{S}^{eff} at 90%. The dashed line plots the corresponding ratios for the STP estimator. Though clearly for both estimators the variances increase as case sampling decreases, the rate of increase is greater for the STP estimator. An STP estimator loses all the information from each missing case; the \tilde{S}^{eff} estimator retains the information contained in $E\left[D_i^{F3} \mid W_i\right]$.

7.3 STP, $\hat{\pi}(\Delta, J)$, and locally efficient nonparametric estimators of survival

The last seven rows of Table 3 contain efficiency results from three *simple local efficient estimators (SLE)*, two *insured local efficient estimators (ILE)*, and the $\hat{\pi}(\Delta, J)$ estimator. Each of the corresponding *SLE's* and *ILE's* use identical estimates, \hat{g}_1^{LE} , of g_1 . However, SLE-estimates are produced by setting $g_1 = \hat{g}_1^{eff}$ in (16, or A3.1), whereas ILE-estimators are $\hat{\pi}$ -estimators based on prediction model (23) with g_1^{eff} replaced by \hat{g}_1^{eff} . By construction, ILE-estimators must be at least as efficient as $\hat{\pi}(\Delta, J)$ -estimators, even when \hat{g}_1^{LE} is based on a misspecified $rr(X_i|v)$ (Mark 2003). SLE's do not share this property. For example, for the *SLE-correct* and *ILE-correct* estimators, \hat{g}_1^{LE} is estimated from a correctly specified model for $rr(X_i|v)$. Specifically, we assumed exponential hazards within each V_i 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 $rr(X_i|v)$'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 $S^s(\tau|v0)$ not only is the SLE-prior estimator less efficient than the $\hat{\pi}(\Delta, J)$ -estimator, it is also 8% less efficient than the STP estimator. In contrast, the ILE-prior, as well as the ILE-null are, in these simulations, as efficient (to two significant digits) as the ILE -correct.

Discussion

Two-stage studies are commonly used in epidemiology as a resource-effective means of estimating the association of disease with exposures whose measurements consume a substrate which is limited in quantity. When estimating survival, the procedures proposed by the case-cohort and nested case-control designs are biased if cases are missing exposure measurements. In this paper, referring to our Linxian studies as examples, we describe how case-missingness arises regardless of investigator intent, and why designs which deliberately sample a fraction of cases are frequently desirable. Applying results of RRZ, we derive a class of nonparametric estimators, and a class of semiparametric estimators, that provide unbiased estimates of cumulative hazards and survivals when cases are missing covariate data. We use a semiparametric estimator to analyze data from a study in which only twenty-five percent of cases were sampled; we find significant differences in age-standardized survivals between subjects with and without serologic evidence of *H. pylori* infection.

Through simulations we demonstrate that the variation in efficiency between estimators within a class is of practical consequence. Efficient estimators make better use of the data observed in stage-one to provide information on the exposures not observed in stage-two. We express the optimal estimators in terms of the familiar quantities of relative risks, survivals, and exposure prevalence; we provide practical strategies for using this formulation to construct estimators with desirable properties. In the design stage, efficiency considerations require collecting information on all covariates suspected of being independent predictors of either exposure or disease. For the analysis stage, we provide a robust procedure ($\hat{\pi}$ -estimation) that incorporates these independent predictors into estimation. S-plus

and R code for implementing these procedures is available from the authors (Mark and Katki, 2003). Given the ease of implementation, and the considerable efficiency advantages even the simplest of $\hat{\pi}$ -estimators possess when compared to the Horvitz-Thompson (STP) estimator, we recommend that the later never be used for estimation of survival from two-stage designs.

Appendix A. Estimating equations, influence functions, and variance estimators

In this appendix random variables are explicitly defined for a univariate counting process such as is appropriate for the semiparametric estimators with CPH model (1). For nonparametric estimators, or semiparametric estimators with a stratified CPH model, the processes should be interpreted in terms of the standard multivariate extension (Anderson et al., 1991). For instance, in nonparametric estimation $\tilde{\Lambda}(\tau, g_1)$ is the $k^* \times 1$ estimator with row entries $\Lambda(\tau, g_1; z)$; $N_i(u)$ is $k^* \times 1$, with the k 'th row defined as $N_{ik}(u) = 1$, iff $I(Z_i = k), T_i \leq u$, and $T_i \leq C_i$. When we can do so without confusion, and to indicate that any consistent estimator of a parameter will suffice, we drop the argument g_b from two-stage estimators and influence functions; e.g., we write $\tilde{\Lambda}(\tau)$ for $\tilde{\Lambda}(\tau, g_1)$. To estimate cumulative hazards and survivals at some time $t \neq \tau$, substitute t for τ in the upper limit of the integrals that define the cumulative hazard estimators.

A.1 Definitions of counting process notation. For more details see Anderson et al. (1991), and for weighted processes, Pugh (1993).

$$N_i(u) = 1 \text{ iff } T_i \leq u, \text{ and } T_i \leq C_i; \quad Y_i(u) = 1, \text{ iff } (C_i \wedge T_i) \leq u.$$

$$dM_i(u) = dN_i(u) - d\Lambda_i(u); \quad d\Lambda_i(u) = Y_i(u) \lambda(u|Z_i).$$

$$S^0(u) = \sum_{j=1}^n R_j Y_j(u); \quad S^0(u, \beta) = \sum_{i=1}^n Y_i(u) \exp(\hat{\beta}^T Z_i).$$

$$d\tilde{M}_i(u) = dN_i(u) - Y_i(u) d\tilde{\Lambda}(u); \quad d\tilde{M}_i(u, \beta) = dN_i(u) - Y_i(u) d\tilde{\Lambda}_o(u, \tilde{\beta}) \exp(\tilde{\beta}^T Z_i).$$

$$\tilde{S}^0(u) = \sum_{j=1}^n \pi_{i,o}^{-1} R_j Y_j(u); \quad \tilde{S}^0(u, \beta) = \sum_{i=1}^n \pi_{i,o}^{-1} R_i Y_i(u) \exp(\tilde{\beta}^T Z_i);$$

$$\tilde{S}^1(u, \beta) = \sum_{i=1}^n Y_i(u) Z_i \exp(\tilde{\beta}^T Z_i); \quad \tilde{E}(u, \beta) = \tilde{S}^1(u, \beta) \tilde{S}^0(u, \beta_o)^{-1}.$$

$$\tilde{i} = n^{-1} \sum_{i=1}^n \pi_{i,o}^{-1} R_i \Delta_i \left(Z_i - \tilde{E}(X_i, \beta) \right) \left(Z_i - \tilde{E}(X_i, \beta) \right)^T; \quad n^{-1} \tilde{S}^j(u, \cdot) \xrightarrow{\text{lim } p} s^j(u, \cdot) \text{ for}$$

$$j \in \{0, 1\}; \quad \tilde{E}(u, \beta) \xrightarrow{\text{lim } p} e(u, \beta_o) = s^1(u, \beta_o) s^0(u, \beta_o)^{-1}; \quad \tilde{i} \xrightarrow{\text{lim } p} i = E \left[\left(\int_0^\tau \{ Z_i - e(u, \beta_o) \} dM_i(u) \right) \left(\int_0^\tau \{ Z_i - e(u, \beta_o) \} dM_i(u) \right)^T \right]. \text{ Here } \xrightarrow{\text{lim } p} \text{ means limit in probability.}$$

A.2 D_i^{Fb} : the full data influence functions (Anderson et al. 1991)

$$D_i^{F1} = \int_0^\tau [s^0(u)]^{-1} dM_i(u); \quad D_i^{F2} = i^{-1} \int_0^\tau \{ Z_i - e(u, \beta_o) \} dM_i(u);$$

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

A.3 Two-stage estimators of $\Lambda(\tau)$, β_o , $\Lambda_o(\tau)$

$$\tilde{\Lambda}(\tau, g_1) = \sum_{i=1}^n \pi_{i,o}^{-1} R_i \int_0^\tau \left(\tilde{S}^0(u) \right)^{-1} dN_i(u) - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_1(W_i). \quad (A3.1)$$

$\tilde{\beta}(g_2)$ is the β that solves

$$\sum_{i=1}^n \int_0^\tau \pi_{i,o}^{-1} R_i \left(Z_i - \tilde{E}(u, \beta) \right) dN_i(u) - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_2(W_i) = 0. \quad (A3.2)$$

To estimate $\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*)$ first estimate $\tilde{\beta}(g_2)$ in A3.2; then

$$\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*) = \sum_{i=1}^n \left\{ \int_0^\tau \pi_{i,o}^{-1} R_i [\tilde{S}^0(u, \tilde{\beta}(g_2))]^{-1} dN_i(u) - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_3^*(W_i) \right\}. \quad (A3.3)$$

A.4 To show that when $b = 3$, (18) is the influence function for $\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*)$, we write

$$\tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*) - \Lambda_o(\tau, \beta_o) = \left\{ \tilde{\Lambda}_o(\tau, \tilde{\beta}(g_2), g_3^*) - \tilde{\Lambda}_o(\tau, \beta_o, g_3^*) \right\} + \left\{ \tilde{\Lambda}_o(\tau, \beta_o, g_3^*) - \Lambda_o(\tau, \beta_o) \right\}.$$

Using a Taylor series expansion of $\tilde{\beta}(g_2)$ around β_o as in Theorem VII 2.3 Anderson et al. (1991), the first term in the right hand side is $(\tilde{\beta}(g_2) - \beta_o)' \int_0^\tau e(u, \beta_o) \lambda_o(u) du + op(1)$. Multiplying by $n^{\frac{1}{2}}$, and replacing estimators with their influence functions gives

$$D_i^3(g_3) = \pi_{i,o}^{-1} R_i D_i^{F3} - \pi_{i,o}^{-1} (R_i - \pi_{i,o}) g_3(W_i); \quad g_{i,3} = g_{i,3}^* - g_{i,2} \int_0^\tau e(u, \beta_o) d\Lambda_o(u, \beta_o). \quad (A4.1)$$

A.5 Estimating $D_i^b(g_b)$ (18), and $D_i^b(\hat{\pi}(W^l))$ (21)

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 $e(u, \beta)$ in the D_i^{Fb} . The weights $\pi_{i,o}$ can be replaced by any consistent estimate, $\hat{\pi}$. For $\hat{\pi}$ -estimation, $\tilde{D}_i^1(\hat{\pi}(W^l))$ and $\tilde{D}_i^2(\hat{\pi}(W^l))$ are formed by estimating P^{bl} (22) 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 . For $b = 3$, the influence function (21) is correct for $\hat{\pi}$ -estimator where $g_2 = 0$. For $\hat{\pi}$ -estimators with any $\tilde{\beta}(g_2)$ used in (A3.3), the influence function is

$$D_i^3(g_2, \hat{\pi}(W^l)) \equiv D_i^3(g_2, g_3^* = 0) - E[D_i^3(g_2, g_3^* = 0) S_i^{l'}] E[S_i^l S_i^{l'}]^{-1} S_i^l. \quad (A5)$$

(A5) is derived by sequential application of RRZ proposition 6.2. The second term in (A5) is estimated by least squares regression as described above. In the particular instance in which the estimates of β_o come from $\hat{\pi}$ -estimation, $g_{i,2}(\pi) = \pi_{i,o} P^{2l} W_i^l$.

A.6 Estimating the asymptotic variance of $\tilde{\Lambda}(\tau)$, and $\{\tilde{\beta}^T \Lambda_o(\tau)\}^T$

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.7 Estimating the asymptotic variances of $\tilde{S}(\tau|vj)$, $\tilde{S}^s(\tau|v)$, $\tilde{R}d(\tau)$.

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 $\tilde{S}(\tau; h, \beta)$; here h is a point in the support of Z_i . 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 $\tilde{V}_{s1} = G\tilde{V}_1G_1$ is a consistent estimate of V_{s1} . Each h 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 $\tilde{V}_{s2} = L\tilde{V}_aL'$ is a consistent estimator of V_{s2} .

Let v^*, j^* be the number of levels in the support of V_i and J_i respectively. Arrange $\tilde{S}(\tau|v, j)$ 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, $C_w \tilde{V}_{s1} C_w^T$. Estimates of standardized risk differences, $\tilde{R}d(\tau)$, are simple contrasts of the $\tilde{S}^s(\tau|v)$. For estimators of population attributable risk and their distribution see (Mark 2003 Appendix A).

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

In this section 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 Δ_i , any $\hat{\pi}$ -estimators with column space greater than (8) can be used.

NCC designs use dependent, risk set, sampling. Let $\{T_{(1)}, \dots, T_{(d)}\}$ be the set of ordered case failure times. We estimate the case-sampling probability, $\pi_{i,o}(\Delta_1)$, 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

$\pi_{i,k} \equiv Pr(R_{ik} = 1 | X_i, \Delta_i = 0, \bar{R}_{ik-1} = 0)$, then

$$Pr(R_i = 1 | \Delta_i = 0, X_i) \equiv \pi_{i,o}(\Delta_0) = \sum_{k=1}^d \pi_{ik} I(X_i \geq T_{(k)}, \bar{R}_{ik-1} = 0) \prod_{j=1}^{k-1} (1 - \pi_{ij}) \quad (B1)$$

where the product term is defined to be 1 when $k = 1$. To estimate $\pi_{i,o}(\Delta_0)$, we replace the π_{ik} in (B1) with the proportion of controls with $(X_i \geq T_{(k)}, \bar{R}_{ik-1} = 0)$ who were sampled at $T_{(k)}$.

Estimating the influence function requires obtaining scores from the likelihood based on $\pi_{i,o}(\Delta)$.

References

Abnet C.C., Borkowf C.B., Qiao Y.L., Albert P.S., Wang E., Merrill A.H., Mark S.D., Dong Z.W., Taylor P.R., Dawsey S.M. (2001). "Sphingolipids as biomarkers of fumonosin exposure and risk of esophageal squamous cell carcinoma," *Cancer Causes and Control*, 12:821-828.

Abnet C.C., Qiao Y.L., Dawsey S.M., Buckman D.W., Yang C.S., Blot W.J., Dong Z.W., Taylor P.R., Mark S.D. (2003). "Prospective Study of Serum Retinol, Beta Carotene, Cryptoxanthin, and Lutein/Zeaxanthin and Esophageal and Gastric Cancers," *Cancer Causes and Control*, 14, 645-655.

Abnet C.C., Lai B., Qiao Y.L., Vogt S., Dong Z.W., Taylor P.R., Mark S.D., Dawsey S.M. (in press), "Zinc concentration in esophageal biopsies measured by X-ray fluorescence and cancer risk," *Journal of the National Cancer Institute*.

Andersen P.K., Borgan O., Gill R.D., and Keiding N. (1993), *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 W.J., Li J.Y. (1985). "Some considerations in the design of a nutritional intervention trial in Linxian, People's Republic of China." *National Cancer Institute Monograph*. 69:29-34.

Blot W.J., Li J.Y., Taylor P.R., Guo W., Dawsey S., Wang G.Q., Yang C.S., Zheng S.F., Gail M., Yu Y. Liu B.Q., Tangera J., Frauweni 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*. 85:1483-92.

- 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.
- Devesa S.S., Blot, W.J., Fraumeni J.F. (1998), "Changing patterns in the incidence of esophageal and gastric carcinoma in the United States," *Cancer*, 83, 2049-2053.
- Horvitz, D.G., and Thompson, D.J. (1952), " A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47, 663-685.
- 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.
- International Agency for Research on Cancer (1994), *Schistosomes, liver flukes and Helicobacter pylori*, Lyon, France.
- 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.
- Mahabir S., Abnet C.C., Qiao Y.L., Ratnasinghe L.D., Dawsey S., Dong Z.W., Taylor P.R., Mark S.D., (Submitted), "Polymorphisms of DNA Repair Genes XRCC1, XPD23, and APE5 and Risk of Stroke in Linxian, China," *Stroke*.
- 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.
- Mark S.D., Selhub J., Qiao Y.L., Buckman D., Dawsey S.M., Blot W.J., Dong Z.W., Taylor P.R. (2001). "Serum cysteine and riboflavin are inversely related to incident esophageal and gastric cardia cancers," *Proceedings, American Association of Cancer Research*, New Orleans LA .
- Mark S.D. (2003). Nonparametric and semiparametric survival estimation in two-stage (Nested) Cohort Studies. *Proceedings of the American Statistical Association, Statistics in Epidemiology Section [CD-ROM]*, 2675-2691. Alexandria, VA: American Statistical Association.
- Mark S.D., and Katki H. (2003). "R and S-PLUS code for $\hat{\pi}$ -estimation of nonparametric and semiparametric estimators of survival and relative risk from two-stage cohort studies," *Technical Report, Biostatistics Branch, Division of Cancer Epidemiology and Genetics*.

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.

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.

Roth M.J., Abnet C.C., Johnson L.L., Mark S.D., Dong Z.W., Taylor P.R., Dawsey A.M., Qiao Y.L. (in press), "Polymorphic variation of CYP1A1 is associated with the risk of gastric cardia cancer: a prospective case-cohort study of phase I and phase II cytochrome P-450 1A1 and GST enzymes," *Cancer Causes and Control*.

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

Samet J.M. and Munoz A. (1998), "Evolution of the cohort Study," *Epidemiologic Reviews*, 20, 1-14.

Savage S.A., Abnet C.C., Haque K., Mark S.D., Qiao Y.L., Dong Z.W., Dawsey S.M., Taylor P.R., Chanock S.J. (in press), "Polymorphisms in Interleukin 2, and Interleukin 10 are not associated with gastric cardia or esophageal cancer in a high-risk Chinese population," *Cancer Epidemiology Biomarkers and Prevention*.

Savage S.A., Abnet C.C., Mark S.D., Qiao Y.L., Dong Z.W., Dawsey S.M., Taylor P.R., Chanock S.J. (in press), "Variants of the IL8 and IL8RB Genes and Risk for Gastric Cardia Adenocarcinoma and Esophageal Squamous Cell," *Cancer Epidemiology Biomarkers and Prevention*.

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.

S-PLUS 6.0 release 2, Insightful Corporation, Seattle, WA.

Stolzenberg-Solomon R., Abnet C.C., Ratnasinghe L., Qiao Y.L., Dawsey S.M., Dong Z.W., Taylor P.R., Mark S.D. (in Press). Esophageal and gastric cardia cancer risks and MTRR A66G and MTHFR C677T and A1298C polymorphisms in Linxian, China. *Cancer Epidemiology Biomarkers and Prevention*.

Taylor P.R., Qiao Y.L., Abnet C.C., Dawsey S.M., Yang C.S., Gunter E.W., Blot W.J., Dong Z.W., Mark S.D. (2003). Prospective study of serum vitamin E levels and esophageal and gastric cancers. Prospective study of serum vitamin E levels and esophageal and gastric cancers," *Journal of the National Cancer Institute*, 95, 1414-1416.

Table 1. Effect of *H. Pylori* infection on age-specific survival and age-standardized survival, at 5.25 years in the Linxian cohort

	<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)
Age Standardized Survival	98.3 (97.7, 98.9)	97.2 (96.7, 97.8)
Age Standardized Risk Difference	1.08 (0.02, 2.15)	

The estimates are based on CPH model (1) with relative risks $\exp(\beta_{01}V_i + \beta_{02}J_i)$

Table 2: Relative efficiencies of the $\hat{\pi}(\Delta)$ and $\hat{\pi}(\Delta, J)$ semiparametric estimators of $S(\tau | v)$ when J is an auxiliary covariate

P(V1 J1)	Relative Efficiency RR _v = 2.0				Relative Efficiency RR _v = 0.5			
	S(τ v0) = 90%		S(τ v1) = 81%		S(τ v0) = 90%		S(τ v1) = 95%	
	$\hat{\pi}(\Delta)$	$\hat{\pi}(\Delta, J)$	$\hat{\pi}(\Delta)$	$\hat{\pi}(\Delta, J)$	$\hat{\pi}(\Delta)$	$\hat{\pi}(\Delta, J)$	$\hat{\pi}(\Delta)$	$\hat{\pi}(\Delta, J)$
.65	82	82	46	46	68	68	62	63
.75	81	79	47	46	67	64	62	61
.85	81	73	47	43	66	58	64	57
.95	80	62	47	40	64	48	65	49

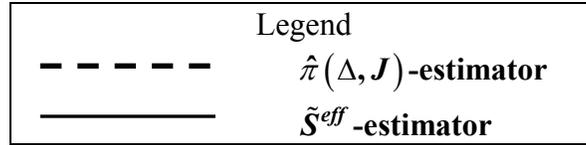
Relative Efficiency equals 100 times the ratio of the variance of an estimator to the variance of the STP estimator. Marginal covariate probabilities are P(V1)=0.65 and P(J1)=0.5.

Table 3:
**Relative Efficiency of RC-Efficient, Locally Efficient and $\hat{\pi}$ -Estimators
of $S^s(\tau | \nu)$ for Semiparametric and Nonparametric Models**

		Relative Efficiency	
	Estimator	$S^s(\tau \nu_0) = 90.4$	$S^s(\tau \nu_1) = 82.0$
Semiparametric	\tilde{S}^{eff}	86	41
	$\hat{\pi}(\Delta, J)$	87	45
Nonparametric	SLE correct	90	42
	ILE correct	90	42
	SLE prior	108	47
	ILE prior	90	42
	SLE null	91	43
	ILE null	90	42
	$\hat{\pi}(\Delta, J)$	90	47

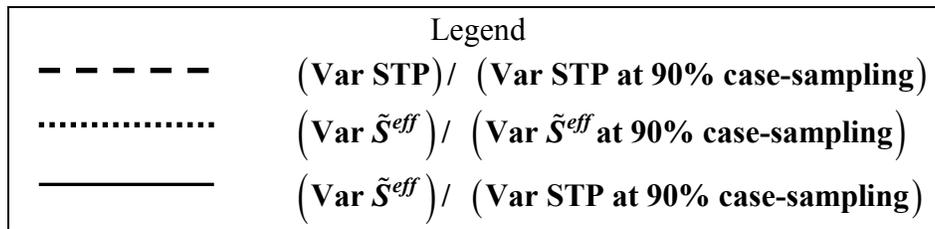
Relative Efficiency equals 100 times the ratio of the variance of the estimator to the variance of the STP estimator. Marginal covariate probabilities are $P(V=1)=0.65$ and $P(J=1)=0.5$, with $P(V=1|J=1)=0.85$ and $P(V=1|J=0)=0.45$

Figure 1. Relative Efficiency of Semiparametric Estimators of $S^s(\tau | \nu_1)$ as Case-sampling Percentage Varies.



The simulation data were generated using the same CPH model as in Table 3. Sampling percent is the binomial sampling probability (x 100) of V measurement. For all simulations control sampling is 15%. Case-sampling percent is indicated on the x-axis. Relative efficiency is defined as 100 times the ratio of the variance of $\hat{\pi}(\Delta, J)$ (dotted line) and RC-efficient (solid line) estimators, to the variance of the STP estimator. Both estimators are substantially more efficient than the STP estimator. The magnitude of the efficiency gains are inversely related to case-sampling percent. Efficiency differences between the \tilde{S}^{eff} and $\hat{\pi}(\Delta, J)$ estimators show a similar dependency on case-sampling percent.

Figure 2. Comparing the Effect of Case-Sampling Percent on the Variance of STP and \tilde{S}^{eff} estimators of $S^s(\tau | \nu_1)$.



The simulation data were generated using the same CPH model as in Table 3. Sampling percent is the binomial sampling probability (x 100) of V measurement. For all simulations control sampling is 15%. Case-sampling percent is indicated on the x-axis. The solid line is the ratio of the variance of the \tilde{S}^{eff} -estimator at each of four case-sampling percents, to the variance of the STP estimator at 90% case sampling. Except at the lowest case-sampling percent, 12.5%, the ratio is less than one. The dotted line compares the variance of the \tilde{S}^{eff} -estimator at each case-sampling percent, to the variance of the \tilde{S}^{eff} -estimator at 90% case-sampling. The dashed line plots the corresponding ratios for the STP estimator. The variances of both estimators increase as case-sampling fractions decrease. The rate of increase is greater for the STP estimator.

Figure 1.

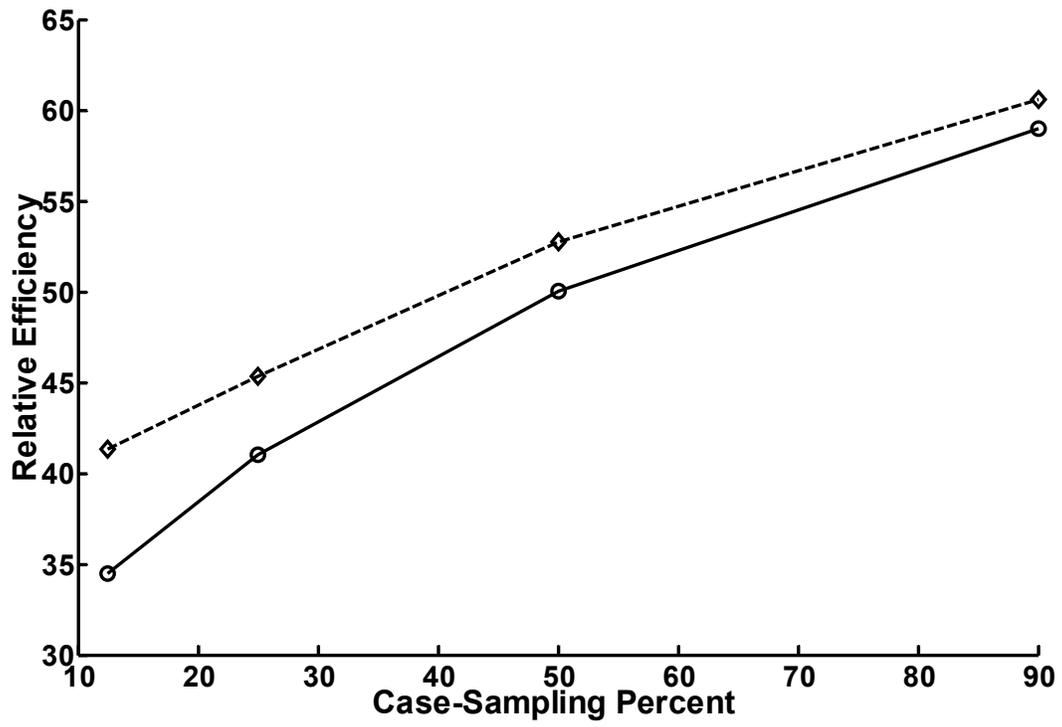


Figure 2.

