University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 36

Nonparametric and semiparametric inference for models of tumor size and metastasis

Debashis Ghosh*

*University of Michigan, debashis.ghosh@ucdenver.edu

http://biostats.bepress.com/umichbiostat/paper36

Nonparametric and semiparametric inference for models of tumor size and metastasis

Debashis Ghosh

Abstract

There has been some recent work in the statistical literature for modelling the relationship between the size of primary cancers and the occurrences of metastases. While nonparametric methods have been proposed for estimation of the tumor size distribution at which metastatic transition occurs, their asymptotic properties have not been studied. In addition, no testing or regression methods are available so that potential confounders and prognostic factors can be adjusted for. We develop a unified approach to nonparametric and semiparametric analysis of modelling tumor size-metastasis data in this article. An equivalence between the models considered by previous authors with survival data structures. Based on this relationship, we develop nonparametric testing procedures and semiparametric regression methodology of modelling the effect of size of tumor on the probability at which metastatic transitions occur in two situations. Asymptotic properties of these estimators are provided. Procedures that achieve the semiparametric information bound are also considered. The proposed methodology is applied to data from a screening study in lung cancer.

Nonparametric and semiparametric inference for models of tumor size and metastasis

D. GHOSH

Department of Biostatistics,

School of Public Health,

University of Michigan,

Ann Arbor, MI, 48109-2029, USA

SUMMARY

There has been some recent work in the statistical literature for modelling the relationship between the size of primary cancers and the occurrences of metastases. While nonparametric methods have been proposed for estimation of the tumor size distribution at which metastatic transition occurs, their asymptotic properties have not been studied. In addition, no testing or regression methods are available so that potential confounders and prognostic factors can be adjusted for. We develop a unified approach to nonparametric and semiparametric analysis of modelling tumor size-metastasis data in this article. An equivalence between the models considered by previous authors with survival data structures. Based on this relationship, we develop nonparametric testing procedures and semiparametric regression methodology of modelling the effect of size of tumor on the probability at which metastatic transitions occur in two situations. Asymptotic properties of these estimators are provided. Procedures that achieve the semiparametric information bound are also considered. The proposed methodology is applied to data from a screening study in lung cancer.

Some key words: Additive risk model; Censoring; Interval censoring; Nonregular Asymptotics; Order-restricted inference; Semiparametric efficiency bound.

1. INTRODUCTION

There has been a rich literature existing on statistical models for tumor progression (Kimmel & Flehinger, 1991; Xu & Prorok, 1997, 1998) in which the phenotype considered was size of the tumor. Solid cancers develop through a process in which tumors originate as a progenitor cell, which grows to a local lesion that shed cancer cells into the lymphatic system and/or blood stream (Foulds, 1969). Some of these cells are transported to distant organs and lead to the development of metastases. In most oncology settings, cancers where metastases have developed are more likely to be associated with worse clinical prognosis. It is thus of vital scientific interest to understand the relationship between tumor size and probability of metastatic spread. This also has implications for the development of screening programs.

Most of the work in this area has focused on nonparametric estimation of the distribution function of tumor size at which metastatic transitions occur. The data that exist are the sizes of tumor samples and an indicator of presence of metastases. Since the data are collected cross-sectionally, the distribution function of tumor size at which metastatic transitions occur is nonidentifiable. Under certain assumptions made by previous authors (Kimmel & Flehinger, 1991; Xu & Prorok, 1997, 1998), this quantity becomes identifiable. In this work, we focus primarily on the proposal of Kimmel & Flehinger (1991), who developed various nonparametric estimation procedures for the distribution function of the tumor size at which metastases occur. However, the asymptotic properties of these estimators have not been studied.

A limitation of the methods described in the previous paragraph is that they do not allow for adjustment of covariates. In the cancer setting, covariates such as the tissue of origin of the tumor or age of the patient can affect the relationship between tumor size and probability of metastatic spread. Thus, it would be desirable to have semiparametric regression models for analyzing such data. However, no such models currently exist.

In this article, we develop a comprehensive approach to the analysis of data on tumor size and metastases. A crucial step in our methodology is the demonstration of the equivalence between the observed data structures with those from survival data analysis. Based on the formulation, we derive asymptotic results for previously proposed estimators in the literature

and formulate hypothesis testing methods and regression generalizations for analyzing data on tumor size and metastases that incorporates other covariates. The structure of the paper is as follows. In §2, we consider the model assumptions about tumor size and progression and relate it to data structures in survival analysis. We present two scenarios in which the distribution function of tumor size at which metastatic transitions occur is identifiable. In §3, we provide asymptotic results for the one-sample distribution function estimators. We also develop hypothesis testing procedures and regression models and estimation procedures for the two scenarios. In §4, the proposed methodology is illustrated with an example from a lung cancer data screening study. Procedures that achieve the semiparametric information bound (Bickel et al., 1993) are outlined in §5. Finally, in §6, we conclude with some brief discussion.

2. NOTATION AND PRELIMINARIES

2.1. Data structure and model assumptions

Let V denote the size of the tumor, Z a p-dimensional vector of covariates and δ be an indicator of tumor metastasis (i.e., $\delta = 1$ if metastases are present, $\delta = 0$ otherwise). We observe the data (V_i, δ_i, Z_i), $i = 1, \ldots, n$, a random sample from (V, δ, Z). In much of the literature previously described in Section 1, only (V_i, δ_i) $(i = 1, \ldots, n)$ were available. We will now state the model assumptions utilized by Kimmel & Flehinger (1991):

1. Primary cancers grow monotonically, and metastases are irreversible.

2. The cancer samples are characterized by the primary tumor sizes at which metastatic transitions take place. We will denote Y as the random variable for this quantity. Let the cdf of Y be denoted by F^Y.

3. Let $\lambda_1(x)$ denote the hazard function for detecting a cancer with metastasis when the tumor size is x. Let $\lambda_0(x)$ denote the hazard function for detecting a cancer with no metastases when the tumor size is x. Assume that $\lambda_1(x) \geq \lambda_0(x)$.

This is also the general framework utilized by Xu & Prorok (1997, 1998). Based on assumptions (1)-(3), F^Y is in general nonidentifiable. However, there are two conditions in

which F^Y becomes identifiable. The first situation is when cancers are detected immediately when the metastasis occurs. The second is when detection of the cancer is not affected by the presence of metastases. We will refer to these situations as Case I and Case II, respectively. Under these two situations, Kimmel & Flehinger (1991) developed nonparametric estimators of F^Y. However, no asymptotic results regarding these estimators were given.

2.2. Equivalences with censored data structures

We now recast the problem in terms of failure time data structures. By assumptions (1) and (3) from the previous section, we can treat Y as a failure time variable. Under the case I scenario (i.e., cancers are detected immediately when the metastasis occurs), we can treat the data structure (V, δ, Z) as a right-censored data structure. For this situation, $V = Y \wedge C$, where C is a random monitoring size, and $\delta = I(Y \leq C)$, where $a \wedge b$ is the minimum of two numbers a and b and $I(A)$ is the indicator function of the event A. Note that under this observation scheme, there is positive probability of Y being observed. What this also means is that the standard methodology for right-censored data (Fleming & Harrington, 1991) can be applied to these data in the Case I situation.

For the Case II scenario, we consider the observed data under the assumption that the detection of the cancer is not affected by the presence of metastases. In this situation, Y is never directly observed. Instead, V is always observed, which represents a monitoring size; thus, (V, δ, Z) has a structure analogous to that of current status data (Groenenboom & Wellner 1992, p. 34). Now the definition of δ is $\delta \equiv I(Y \leq U)$. Note that the definition of δ will change depending on whether we are talking about the Case I or Case II scenarios; the appropriate choice of δ should be evident from the context. Because Y is never directly observed for the Case II situation, there is inherently less information available about the distribution of Y than in the Case I scenario. This will affect the asymptotic results for the nonparametric estimators of tumor size at which metastatic distributions occur.

3. STATISTICAL METHODOLOGY

In this section, we describe the appropriate statistical methodology for the Case I and II scenarios. By §2.2, the tumor size for these situations can be treated as survival times, so

existing methodologies for the analysis of survival data can be applied to the tumor size. In §3.1 and §3.2, we develop nonparametric and semiparametric procedures for modelling tumor size-metastasis data in the Case I and Case II scenarios.

Before proceeding, we make two comments. The first is regarding regression models. The standard regression model for the analysis of failure time data is the proportional hazards model (Cox, 1972):

$$\lambda(y|Z) = \lambda_0(y)\exp(\alpha_0' Z), \tag{3.1}$$

where $\lambda(y|Z)$ is the conditional hazard function of Y given Z, α_0 is a p-dimensional vector of unknown regression coefficients, and $\lambda_0(y)$ is an unspecified baseline hazard function. In model (3.1), α has a relative risk interpretation on a logarithmic scale, and estimation of α has been well-studied for both right-censored data (Cox, 1972; Andersen & Gill, 1982) and for interval-censored data (Huang, 1996). However, as Kimmel & Flehinger (1991) write, for studying the relationship of tumor size on the hazard function of Y, the additive risk model (Breslow & Day, 1980, p. 53 – 57) is more applicable: "we prefer the additive model because we associate it with a constant hazard of detection of metastases that does not depend on the size of the primary tumor." In this article, we formulate regression models for the hazard function of Y using the additive risk model, i.e.

$$\lambda(y|Z) = \lambda_0(y) + \beta_0' Z, \tag{3.2}$$

where β_0 is a p-dimensional vector of unknown regression coefficients. We comment on the use of the proportional hazards model (3.1) in the discussion.

The second comment deals with the issue of nonmeasured tumors. In the framework considered by Kimmel & Flehinger (1991), there was the possibility that there were tumors, both with and without metastases, for which the tumor size was not measured. In the framework presented in §2.2, this corresponds to observations with missing observed failure time measurements. In this article, we assume that the proportion of nonmeasured tumors relative to the total number of tumors is $o_P(n^{-1/2})$. This will imply that the limiting distributions of the estimators proposed here with those proposed by Kimmel & Flehinger (1991) will be equivalent. In order to develop estimation procedures with nonmeasured tumors, some type

of imputation method (Little & Rubin, 2002) would be required.

3.1. Procedures for the Case I scenario

We first consider the case I situation. In this case, we can treat V as a right-censored version of Y. For nonparametric estimation of the survival function corresponding to F^Y, say S^Y, the estimator of Kaplan & Meier (1958) can be used. Let $v_{(1)} < v_{(2)} < \cdots < v_{(m)}$ denote the sorted tumor sizes. Then the Kaplan-Meier estimator is given by

$$\hat{S}^Y(y) = \Pi_{i:v_{(i)}<y}\left(1 - \frac{d_i}{n_i}\right),$$

where d_i is the number of tumors with metastases with size $v_{(i)}$, and n_i is the number of tumors of size at least $v_{(i)}$, $i = 1, \ldots, m$. An alternative estimator of S^Y can be derived using the exponentiated Nelson-Aalen type estimator

$$\tilde{S}^Y(y) = \exp\left(-\sum_{i:v_{(i)}<y}\frac{d_i}{n_i}\right).$$

Asymptotically, the difference between $\hat{S}^Y$ and $\tilde{S}^Y$ will be negligible. The estimator $\hat{S}^Y$ is the estimator proposed by Kimmel & Flehinger (1991) and is asymptotically equivalent to the estimator of Xu & Prorok (1997) in the case where $C = 1$ (in their notation).

Before proving asymptotic results about the estimator $\hat{S}^Y$, we introduce some notation. Let $N_i(t) = I(V_i \leq t, \delta_i = 1)$ and $R_i(t) = I(V_i \geq t)$, $i = 1, \ldots, n$. The following result can be proven by standard survival analysis techniques (Fleming & Harrington, 1991).

Theorem 1: *Assuming the usual regularity conditions, $n^{1/2}(\hat{S}^Y - S^Y)$ converges weakly to a mean-zero Gaussian process with covariance function*

$$\xi(s,t) = S^Y(s)S^Y(t)\int_0^{s\wedge t}\frac{d\Lambda(u)}{\pi(u)},$$

where $\Lambda(t)$ is the cumulative hazard function corresponding to F_Y, and

$$\pi(t) = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} R_i(t).$$

The covariance function in Theorem 1 can be estimated consistently using empirical quantities. A similar result to Theorem 1 can be proven for the limiting distribution of $n^{1/2}(\tilde{S}^Y - S^Y)$.

Suppose that $p = 1$ and that Z is a discrete covariate taking values $(1, \ldots, K)$, where $K \geq 1$. The data can be expressed as $\{V_{ij}, \delta_{ij}, i\}$, $j = 1, \ldots, n_i$ and $i = 1, \ldots, K$. We define $N_{ij}(t) = I(V_{ij} \leq t, \delta_{ij} = 1)$, $R_{ij}(t) = I(V_{ij} \geq t)$, $R_{i\cdot}(t) = \sum_{j=1}^{n} R_{ij}(t)$, $N_{i\cdot}(t) = \sum_{j=1}^{n} N_{ij}(t)$, $R_{\cdot\cdot}(t) = \sum_{i=1}^{n} R_{i\cdot}(t)$, and $N_{\cdot\cdot}(t) = \sum_{i=1}^{n} N_{i\cdot}(t)$. In order to test $H_0 : F_0^Y = F_1^Y = \cdots F_K^Y$, we can utilize the G^ρ family of test statistics proposed by Harrington & Fleming (1982):

$$T = Z'\Sigma Z,$$

where $Z = \{Z_1(\tau), \ldots, Z_K(\tau)\}$, $\tau > 0$ is a truncation time assumed to satisfy certain technical conditions,

$$Z_i(t) = \int_0^t K(s) dN_{i\cdot}(s) - \int_0^t K(s) \frac{R_{i\cdot}(s)}{R_{\cdot\cdot}(s)} dN_{\cdot\cdot}(s),$$

$K(t) = \{\hat{S}^Y(t)\}^\rho I\{R_{\cdot\cdot}(s) > 0\}$ and Σ is a $K \times K$ matrix with (l, m)th element

$$\sigma_{lm} = \int_0^t K^2(s) \frac{R_{l\cdot}(s)}{R_{\cdot\cdot}(s)} \left(\delta_{lm} - \frac{R_{m\cdot}(s)}{R_{\cdot\cdot}(s)} \right) dN_{\cdot\cdot}(s),$$

and $\delta_{lm} = I(Z_l = Z_m = l)$. Using arguments in §5.2 of Andersen et al. (1993), T converges in distribution to a χ^2 random variable with $K - 1$ degrees of freedom. The choice of ρ will affect the power of the test and will depend on what type of alternatives one wishes to have high probability of detecting.

Finally, we can formulate a semiparametric model for the effect of covariates on the hazard function through equation (3.2). Estimation in this model has been previously developed by Lin & Ying (1994). The following estimating function can be used for estimation of β in (3.2):

$$U(\beta) = \sum_{i=1}^{n} \int_0^\infty \{Z_i - \bar{Z}(t)\}\{dN_i(t) - R_i(t)\beta' Z_i dt\}, \tag{3.3}$$

where $\bar{Z}(t) = \sum_{j=1}^{n} R_j(t) Z_j / \sum_{j=1}^{n} R_j(t)$. As noted by Lin & Ying (1994), setting $U(\beta)$ from (3.3) equal to zero yields a closed-form estimator for β_0:

$$\hat{\beta} = \left[\sum_{i=1}^{n} \int_0^\infty R_i(t) \{Z_i - \bar{Z}(t)\}^{\otimes 2} dt \right]^{-1} \left[\sum_{i=1}^{n} \int_0^\infty \{Z_i - \bar{Z}(t)\} dN_i(t) \right],$$

where $a^{\otimes 2} = aa'$. By standard martingale arguments (Lin & Ying, 1994), the random vector $n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution to a p-dimensional normal random vector with mean zero and variance $A^{-1}BA^{-1}$, where

$$A = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \int_0^\infty R_i(t)\{Z_i - \bar{Z}(t)\}^{\otimes 2} dt$$

and

$$B = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \int_0^\infty \{Z_i(t) - \bar{Z}(t)\}^{\otimes 2} dN_i(t).$$

The procedure for estimating the regression coefficients does not achieve the semiparametric efficiency bound. In §5, we address this issue.

3.2. Procedures for the Case II scenario

We now deal with the situation where Y is never directly observed and the observed data structure mimics that found with interval-censored data. The one-sample problem is first considered. A precise characterization of F^Y in this situation is found in Groenenboom & Wellner (1992, pp. 38 – 40). Let $v_{(1)} \leq v_{(2)} \leq \cdots \leq v_{(n)}$ denote the observed order statistics for $(V_1, \ldots, V_n)$, and let $\delta_{(i)}$ $(i = 1, \ldots, n)$ denote the corresponding value of δ. Define $v_{(0)} = 0$ and $\delta_{(0)} = 0$. The nonparametric maximum likelihood estimator (NPMLE) of F^Y corresponds to the point $\tilde{x} \equiv (\tilde{x}_1, \ldots, \tilde{x}_n)$ that maximizes

$$f(x_1, \ldots, x_n) = \sum_{i=1}^{n} \{\delta_{(i)} \log x_i + (1 - \delta_{(i)}) \log(1 - x_i)\}$$

over $(x_1, \ldots, x_n) \in R^n$ subject to the constraint

$$0 \leq x_1 \leq \cdots \leq x_n \leq 1.$$

We derive the NPMLE of F^Y, $\hat{F}^Y_*$, through the relationship $\tilde{x}_i = \hat{F}^Y_*(v_{(i)})$, $i = 0, \ldots, n$. Note that the NPMLE of F^Y is defined only up to the set of observed times. The solution to this optimization problem can be characterized in one of two ways. The first is using the so-called "max-min formula" (Groenenboom & Wellner, 1992, p. 40):

$$\tilde{x}_m = \max_{i\leq m} \min_{k\geq m} \frac{\sum_{i\leq j\leq k} \delta_{(j)}}{k - i + 1},$$

$m = 0, \ldots, n$. A second representation of the maximizer is more graphical in nature. One plots the points $\{i, \sum_{j \le i} \delta_{(j)}\}$ $(i = 0, \ldots, n)$ and draws the greatest convex minorant of these points, defined as the function H^* such that

$$H^*(t) = \sup\{H(t) : H(i) \le \sum_{j \le i} \delta_{(j)}, H(0) = 0, H \text{ is convex } \}.$$

Then $\tilde{x}_i$ is the left derivative of H^* at $i = 0, \ldots, n$. This estimator corresponds to that proposed by Kimmel & Flehinger (1991) in the Case II scenario. They provided no asymptotic analysis of this estimator. Using the arguments in Chapter 5 of Groenenboom & Wellner (1992), we can prove the following result:

Theorem 2: *Define* $G(t) = Pr(V \le t)$. *Let* z_0 *be such that* $0 < F^Y(z_0) < 1$ *and* $0 < G(z_0) < 1$. *Assume that* F^Y *and* G *are differentiable at* z_0 *with strictly positive derivatives* $f^Y(z_0)$ *and* $g(z_0)$, *respectively. Then* $n^{1/3}\{\hat{F}^Y_*(z_0) - F^Y(z_0)\}$ *converges in distribution to the random variable* CZ, *where*

$$C = \left[\frac{4F^Y(z_0)\{1 - F^Y(z_0)\}f^Y(z_0)}{g(z_0)}\right]^{1/3}$$

and $Z \equiv argmin\{W(t) + t^2\}$, *and* W *is two-sided Brownian motion starting from zero.*

Note that the limiting distribution presented in Theorem 2 is much different than that in Theorem 1. This is because in the case II scenario, Y is never directly observed. This leads to the slower convergence rate and the more complicated limiting distribution.

A 95% confidence interval for $F^Y(z_0)$ is then given by

$$\{\hat{F}^Y(z_0) - n^{-1/3}\,\hat{Q}_{.975}, \hat{F}^Y(z_0) - n^{-1/3}\,\hat{Q}_{.975}\},$$

where $\hat{Q}_{.975}$ is a consistent estimator of $Q_{.975}$, the 97.5th percentile of the limiting random variable CZ. But $Q_{.975}$ is simply $C \times .99818$ where .99818 is the 97.5th percentile of Z, where the quantiles of Z are from Groenenboom and Wellner (2001). Since C involves the unknown parameters $G(z_0)$, $h(z_0)$, and $g(z_0)$, we estimate C by

$$\widehat{C}_n = \left[\frac{4\hat{f}^Y(z_0)\hat{F}^Y(z_0)\{1 - \hat{F}^Y(z_0)\}}{\hat{g}(z_0)}\right]^{1/3},$$

where $\hat{f}^Y$ and $\hat{g}$ are estimates of f^Y and g. An asymptotic 95% confidence interval is then given by

$$\left\{\hat{F}^Y(z_0) - n^{-1/3}\,\widehat{C}_n \times .99818\ ,\ \hat{F}^Y(z_0) + n^{-1/3}\,\widehat{C}_n \times .99818\right\}.$$

Based on Theorem 2, constructing confidence intervals for $F^Y(z_0)$ requires consistent estimation of f^Y and g. While g can be estimated consistently using nonparametric regression methods, it is much more difficult to estimate f^Y because Y is never directly observed. We estimate g using kernel density methods, while f^Y is estimated using a numerical derivative based on a smoothing spline-based estimate of F^Y (Heckman & Ramsay, 2000).

Proceeding as in the Case I situation, we now consider the problem of testing the equality of the distribution function of Y across K groups. A simple test in this situation is found by modifying the method of Sun (2000). For simplicity, assume that the distribution of S is equal across the K groups. In this case, Z is $K-1$ dimensional. Define the counting process $W_i(t) \equiv I(Y_i \leq t)$. A test of the null hypothesis $H_0 : F_0^Y = F_1^Y = \cdots F_K^Y$ is given by the statistic

$$\tilde{T} = \sum_{i=1}^{n} (Z_i - \bar{Z}) W_i(V_i),$$

where $\bar{Z} = n^{-1} \sum_{j=1}^{n} Z_j$. Under the null hypothesis, Sun (2000) shows that $n^{-1/2}\tilde{T}$ has a limiting normal distribution with mean zero and covariance matrix which is consistently estimated by $n^{-1} \sum_{i=1}^{n} (Z_i - \bar{Z})^2 W_i^2(V_i)$.

We now consider estimation in the regression formulation (3.2) in the Case II scenario. Lin, Oakes & Ying (1998) proposed a procedure for estimation of β_0 in model (2.1) based on the partial likelihood using the counting process $\tilde{N}_{1i}(t) \equiv (1-\delta_i) I(S_i \leq t)$, $i = 1, \ldots, n$. Let $dH_i(t)$ denote this hazard corresponding to $d\tilde{N}_{1i}(t)$, the increment in $\tilde{N}_{1i}(t)$, $i = 1, \ldots, n$. Lin et al. (1998) show that under model (3.2), the following model for $dH_i(t)$ $(i = 1, \ldots, n)$ is induced:

$$dH_i(t) = dH_0(t) \exp\{\beta_0' Z_i^*(t)\}, \tag{3.4}$$

where $dH_0(t) = \exp\{-\Lambda_0(t)\} d\Lambda^S(t)$, $\Lambda^S(t)$ is the cumulative hazard function of S, $\Lambda_0(t) = \int_0^t \lambda_0(u) du$, and $Z_i^*(t) = -tZ_i$. The model in (3.4) has a form identical to that of the proportional hazards model (Cox, 1972). We can estimate β_0 using the partial likelihood score function:

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i^*(t) - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} d\tilde{N}_{1i}(t), \tag{3.5}$$

where $S^{(k)}(\beta, t) = n^{-1} \sum_{j=1}^{n} R_j(t) Z_j^*(t)^{\otimes k} \exp\{\beta' Z_j^*(t)\}, k = 0, 1, 2$, $a^{\otimes 0} = 1$ and $a^{\otimes 1} = a$.

The constant $\tau > 0$ is a truncation time chosen to satisfy certain technical conditions; in practice, we can choose it to be the largest monitoring time. Let $\widehat{\beta}$ be the solution from setting $U(\beta)$ in (3.5) equal to zero. Using martingale theory, $n^{1/2}(\widehat{\beta} - \beta_0)$ converges in distribution to a normal random vector with mean zero and variance $\mathcal{I}(\beta_0) \equiv \lim_{n\to\infty} n^{-1} I(\beta_0)$, where

$$I(\beta) = \sum_{i=1}^{n} \int_0^\tau \left\{ \frac{S^{(2)}(\beta,t)}{S^{(0)}(\beta,t)} - \frac{S^{(1)}(\beta,t)^{\otimes 2}}{S^{(0)}(\beta,t)^2} \right\} d\tilde{N}_{1i}(t).$$

This variance can be consistently estimated based on empirical quantities.

The primary advantage of the Lin, Oakes & Ying (1998) method for estimation in the additive risk model is its simplicity. However, this procedure does not achieve the semiparametric information bound. We discuss efficient estimation procedures for the Case II scenario in §6.

4. NUMERICAL EXAMPLE

In this section, we consider data from a screening trial involving lung cancer and reported in Kimmel & Flehinger (1991). The lung cancer data was collected on a population of male smokers over 45 years old enrolled in a clinical trial involving sputum cytology. There are two types of lung cancer diagnosed, adenocarcinomas (cancers that originate in epithelial cells) and epidermoid cancer (cancers that originate in the epidermis). For the adenocarcinomas, they were detected by radiologic screening and by symptoms; the epidermoids were detected by sputum cytology or by chest X-ray. Presence or absence of metastasis was determined using available staging, clinical, surgical and pathological readings. There are 141 adenocarcinomas, of which 19 have metastases; of the 87 epidermoid cancers, 6 have metastases. The proposed techniques from §3 are now applied.

First, the tumor size and metastases data are analyzed under the Case I situation, i.e. metastases occur at the time of detection. In this instance, we can treat the tumor size as a right-censored variable. The estimated survival functions for the size distribution and pointwise 95% confidence intervals for the adenocarcinomas and for the epidermoid cancers are given in Figures 1 and 2, respectively. Next, differences in the size distribution between the two types of tumors was tested using the Fleming-Harrington G^ρ class of statistics.

Results for various values of ρ are given in Table 1. We find that there is slight evidence of a difference in size distributions between the two types of lung tumors. Finally, we analyzed the data using additive risk model (3.2) in which there is one covariate Z, a binary indicator for tumor site (0/1 = adenocarcinoma/epidermoid). The estimated regression coefficient was 0.0215, with an associated standard error of 0.0150. Based on the Wald statistic, we have a p-value of 0.15, which again suggests a marginal association.

Next, we analyzed the data using the Case II scenario; here, the tumor size is now an interval censored random variable. First, we plot the survival functions for the tumor size distributions and associated pointwise 95% confidence intervals; these graphs for the adenocarcinoma and epidermoid lung cancers are given in Figures 3 and 4, respectively. In testing for a difference in tumor size distributions between tumor type (adenocarcinoma versus epidermoid), the method of Sun (2000) yields a test statistic of 1.60 and an associated p-value of 0.11. Finally, the estimation procedure for the semiparametric additive hazards model of Lin et al. (1998) with Z as a binary indicator for tumor site yields an estimated regression coefficient of 0.16 and standard error of 0.10. The Wald statistic gives a corresponding p-value of 0.11, which suggests a strong association between tumor type and hazard of tumor size detection.

5. EFFICIENT ESTIMATION

In this section, we focus on semiparametric efficient procedures for the additive hazards model in the Case I and II scenarios. We first consider the case I scenario. Then results from Lin & Ying (1994) can be utilized here. We start by considering the following generalization of (3.2):

$$\lambda(t|Z) = \lambda_0(t) + \theta'\mu(t) + \beta'Z \qquad (6.1),$$

where θ is a p-dimensional vector of unknown regression coefficients and $\mu(t)$ is a p-dimensional function of time. In order to determine the semiparametric information bound for β at β_0, we need to calculate the supremum among all parametric information bounds for β in (6.1) over all possible choices of $\mu(t)$ when $\theta = 0$ and $\beta = \beta_0$ (Bickel et al., 1993). Following the derivation of Lin and Ying (1994), the least favorable submodel of (6.1) occurs when

$\mu(t) = \mu_0(t)$, where

$$\mu_0(t) = \lim_{n\to\infty} \frac{\sum_{i=1}^n E[R_i(t)Z_i/\{\lambda_0(t) + \beta_0' Z_i\}]}{\sum_{i=1}^n E[R_i(t)/\{\lambda_0(t) + \beta_0' Z_i\}]}.$$

This model yields the semiparametric information bound

$$I_{sp} \equiv \left(\lim_{n\to\infty} \int_0^\infty n^{-1} \sum_{i=1}^n E\left[\frac{R_i(t)\{Z_i - \mu_0(t)\}^{\otimes 2}}{\lambda_0(t) + \beta_0' Z_i} \right] dt \right)^{-1}.$$

We can then derive the optimal estimating function for β_0 whose variance asymptotically achieves the semiparametric information bound:

$$U_{opt}(\beta) = \sum_{i=1}^n \int_0^\infty \frac{\{Z_i - \bar{Z}(t)\}}{\lambda_0(t) + \beta_0' Z_i} \{dN_i(t) - R_i(t)\beta' Z_i dt\}, \tag{6.2}$$

where $\bar{Z} = \sum_{j=1}^n R_j(t)Z_j / \sum_{j=1}^n R_j(t)$. Note that $U_{opt}(\beta)$ in (6.2) depends on $\lambda_0(t)$, which is unknown. In order to implement $U_{opt}(\beta)$ in practice, one would need to use sample splitting methods. Further details on numerically implementing the estimating function $U_{opt}(\beta)$ can be found in Lin and Ying (1994).

We move now the situation corresponding to the Case II situation. In this setting, semiparametric efficient procedures have been provided by Ghosh (2001) and Martinussen and Scheike (2002); here, we adapt results given by the latter. We will derive the efficient score function to calculate the semiparametric information bound. The log-likelihood function is proportional to

$$l(\beta, \Lambda) = \sum_{i=1}^n \Delta_i \log\left(e^{-\Lambda(S_i) - \beta' Z_i} \right) + (1 - \Delta_i) \log\left(1 - e^{-\Lambda(S_i) - \beta' Z_i} \right).$$

Define $\tilde{N}_{2i}(t) = \delta_i I(Y_i \le t)$ $(i = 1, \ldots, n)$. By the arguments in the Appendix of Martinussen and Scheike (2002), the efficient score for β is

$$\dot{l}^*_\beta = \int \left\{ Z_1 - E\left(Z_1 R_1(t) \frac{w}{1-w} \right) E\left(R_1(t) \frac{w}{1-w} \right)^{-1} \right\} \left(\frac{w}{1-w} d\tilde{N}_{21}(t) - d\tilde{N}_{11}(t) \right),$$

where $w = w(t, \Lambda, \beta) = \exp\{-\Lambda(t) - \beta' Z\}$. The semiparametric information bound is then given by $\tilde{I}_{sb} \equiv E[\dot{l}^*_\beta (\dot{l}^*_\beta)']$. In practice, one uses the empirical score equation to estimate β, i.e. solve $U(\beta, \hat{\Lambda}) = 0$, where $\hat{\Lambda}$ is an estimator of Λ,

$$U(\beta, \Lambda) = \sum_{i=1}^n \int \left(Z_i - \frac{\tilde{S}^{(1)}(t)}{\tilde{S}^{(0)}(t)} \right) \left(\frac{w_i(t)}{1 - w_i(t)} d\tilde{N}_{2i}(t) - d\tilde{N}_{1i}(t) \right),$$

and $\tilde{S}^{(k)}(t) = n^{-1}\sum_{j=1}^{n} w_i(1-w_i)^{-1}R_i(t)Z_i(t)^{\otimes k}$. In practice, some type of smoothing will be required for implementing the efficient estimation procedure. This was utilized by Martinussen and Scheike (2002).

6. DISCUSSION

In this article, we have laid out a general framework for the analysis of data on tumor size and metastases with covariates. The key step in the development of procedures was the equivalence of the observed data structure with those from the field of survival analysis. This relationship allowed us to characterize nonparametric maximum likelihood estimators of the distribution function for tumor size at which metastatic transitions occur and their associated asymptotic properties. In addition, we have been able to develop testing and regression procedures with such data.

In terms of regression methodologies, we have primarily focused on the additive hazards model. However, extension to the proportional hazards model for the Case I and II situations is straightforward by applying the methods of Cox (1972) and Huang (1996).

Much of the literature in the area of cancer screening has focused on mechanistic models for tumor progression (Yakovlev & Tsodikov, 1996). Such models tend to be parametric in nature, while the methods we have proposed here are less parametric. These procedures could serve as an exploratory device in order to determine what parametric models can be utilized.

ACKNOWLEDGMENTS

The author thanks Jeremy Taylor, Moulinath Banerjee and Pinaki Biswas for helpful discussions.

REFERENCES

ANDERSEN, P. K., BORGAN, Ø., GILL, R. D. & KEIDING, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

BICKEL, P. J., KLAASSEN, C. A. J., RITOV, Y. & WELLNER, J. A. W. (1993). *Efficient and Adaptive Estimation for Semiparametric Models.* Baltimore: Johns Hopkins University Press.

BRESLOW, N. & DAY, N. E. (1980). *Statistical Methods in Cancer Research (Vol. 1): The Analysis of Case-control Studies.* Lyon: IARC..

COX, D. R. (1972). Regression models and life tables (with discussion). *J. R. Statist. Soc. B*, **34**, 187–220.

FLEMING, T. R. & HARRINGTON, D. P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

FOULDS, L. (1969). Characteristics of neoplasms. In *Neoplastic Development*, Volume 1, Ed. L. Foulds , pp. 97 – 136. London: Academic Press.

GHOSH, D. (2001). Efficiency considerations in the additive hazards model with current status data. *Stat. Neerl.* **55**, 367 – 376.

GROENENBOOM, P. & WELLNER, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation.* Basel: Birkhäuser.

HARRINGTON, D. P. & FLEMING, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 133 – 143.

HECKMAN, N. E. & RAMSAY, J. O. (2000) Penalized regression with model-based penalties. *Canadian Journal of Statistics* **28**, 241 – 258.

HUANG, J. (1996). Efficient estimation for the proportional hazards model with interval censoring. *Ann. Statist.* **24**, 540–68.

KIMMEL, M. & FLEHINGER, B. J. (1991). Nonparametric estimation of the size-metastasis relationship in solid cancers. *Biometrics* **47**, 987 – 1004.

LIN, D. Y., OAKES, D. & YING, Z. (1998). Additive hazards regression with current status data. *Biometrika* **85**, 289–298.

LIN, D. Y. & YING, Z. (1994). Semiparametric analysis of the additive risk model. *Biometrika* **81**, 61–71.

LITTLE, R. J. A. & RUBIN, D. B. (2002). *Statistical Analysis with Missing Data, 2nd Edition.* New York: Wiley.

MARTINUSSEN, T. & SCHEIKE, T. H. (2002). Efficient estimation in additive hazards regression with current status data. *Biometrika* **89**, 649 – 658.

SUN, J. (2000). A nonparametric test for current status data with unequal censoring. *JRSS-B* **61**, 243 – 250.

XU, J. L. & PROROK, P. C. (1997). Nonparametric estimation of solid cancer size at metastasis and probability of presenting with metastasis at detection. *Biometrics* **53**, 579 – 591.

XU, J. L. & PROROK, P. C. (1998). Estimating a distribution function of the tumor size at metastasis. *Biometrics* **54**, 859 – 864.

YAKOLEV, A. Y. & TSODIKOV, A. D. (1996). *Stochastic Models of Tumor Latency and Their Biostatistical Applications.* Singapore: World Scientific Press.

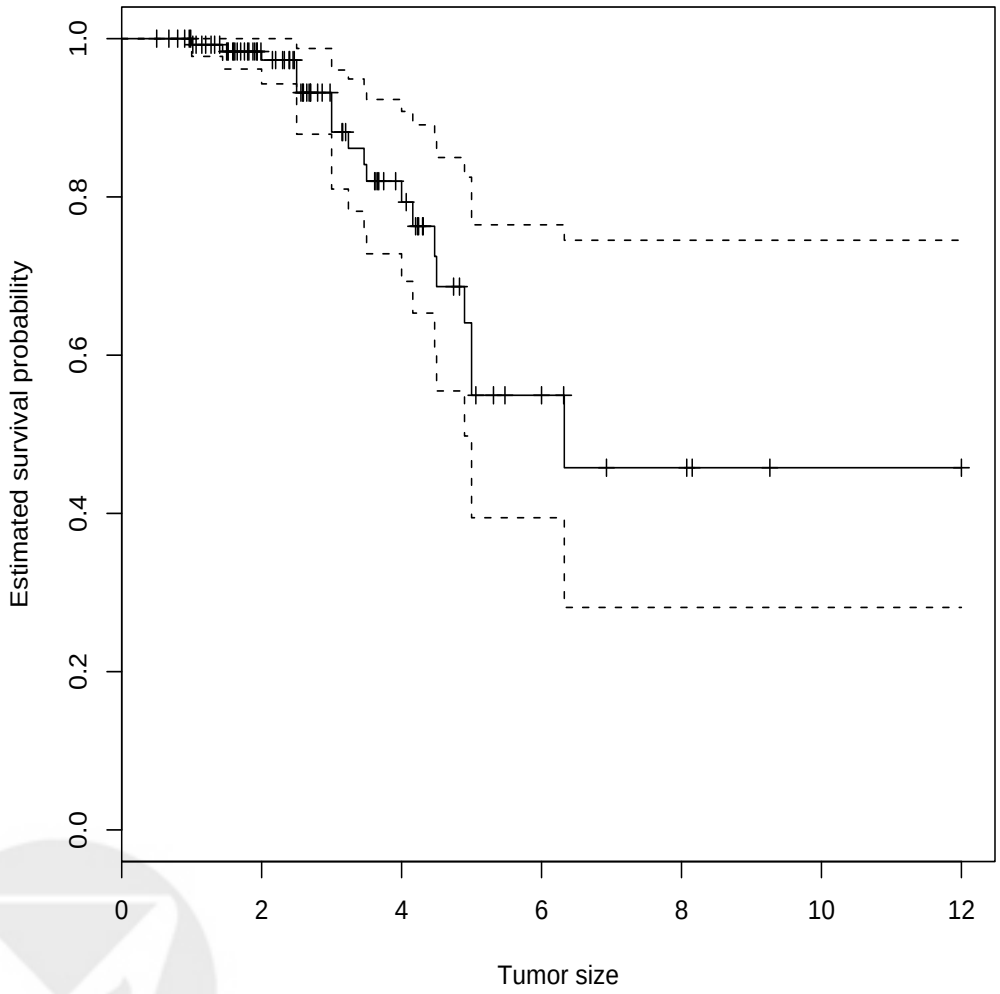

Figure 1: Distribution of survival function for tumor size under Case I scenario for lung adenocarcinoma data (solid line) and 95% pointwise confidence intervals (dashed lines).

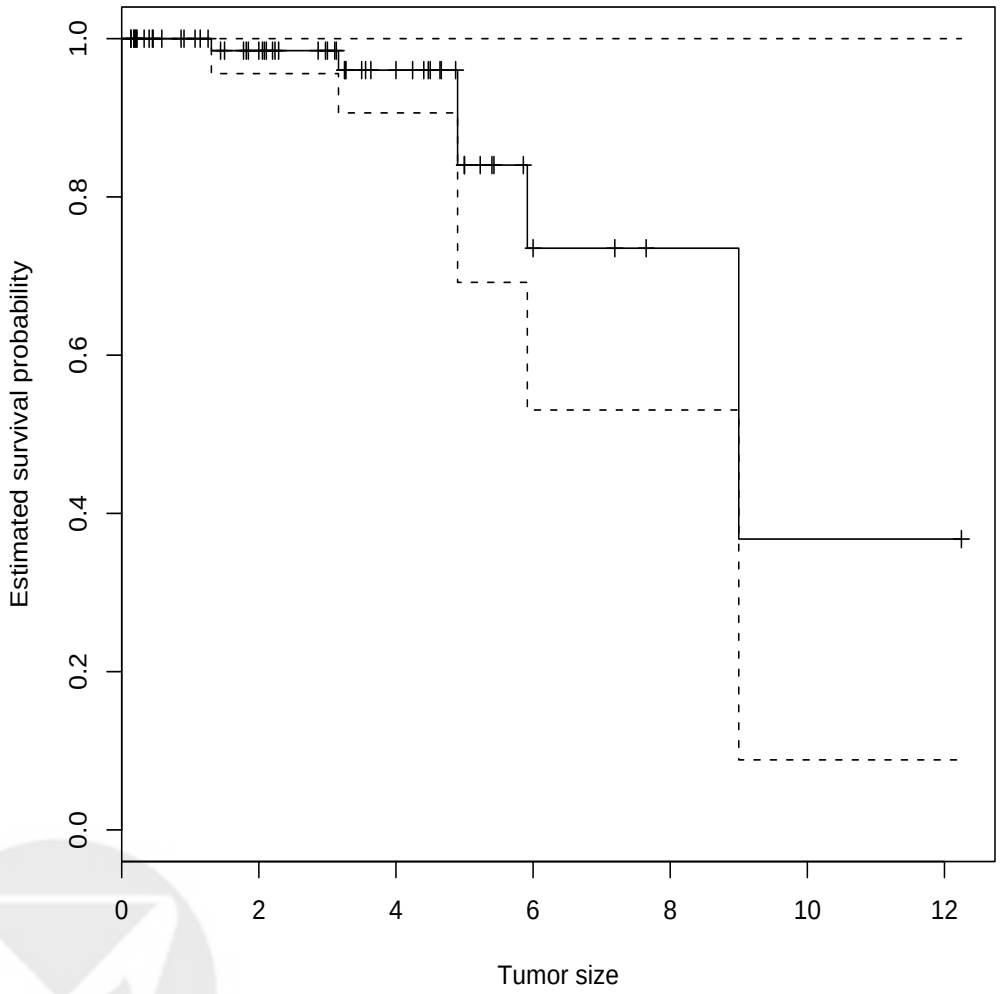

Figure 2: Distribution of survival function for tumor size under Case I scenario for lung epidermoid data (solid line) and 95% pointwise confidence intervals (dashed lines).

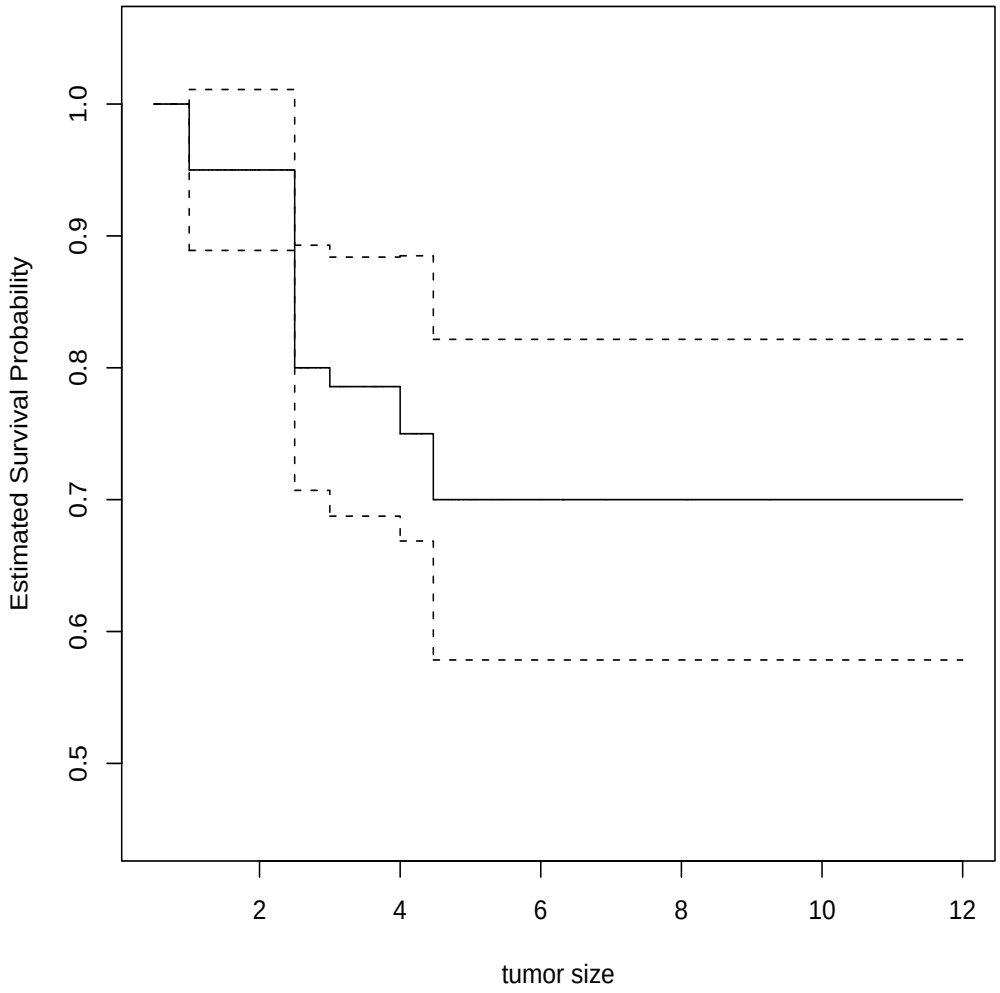

Figure 3: Distribution of survival function for tumor size under Case II scenario for lung adenocarcinoma data (solid line) and 95% pointwise confidence intervals (dashed lines).

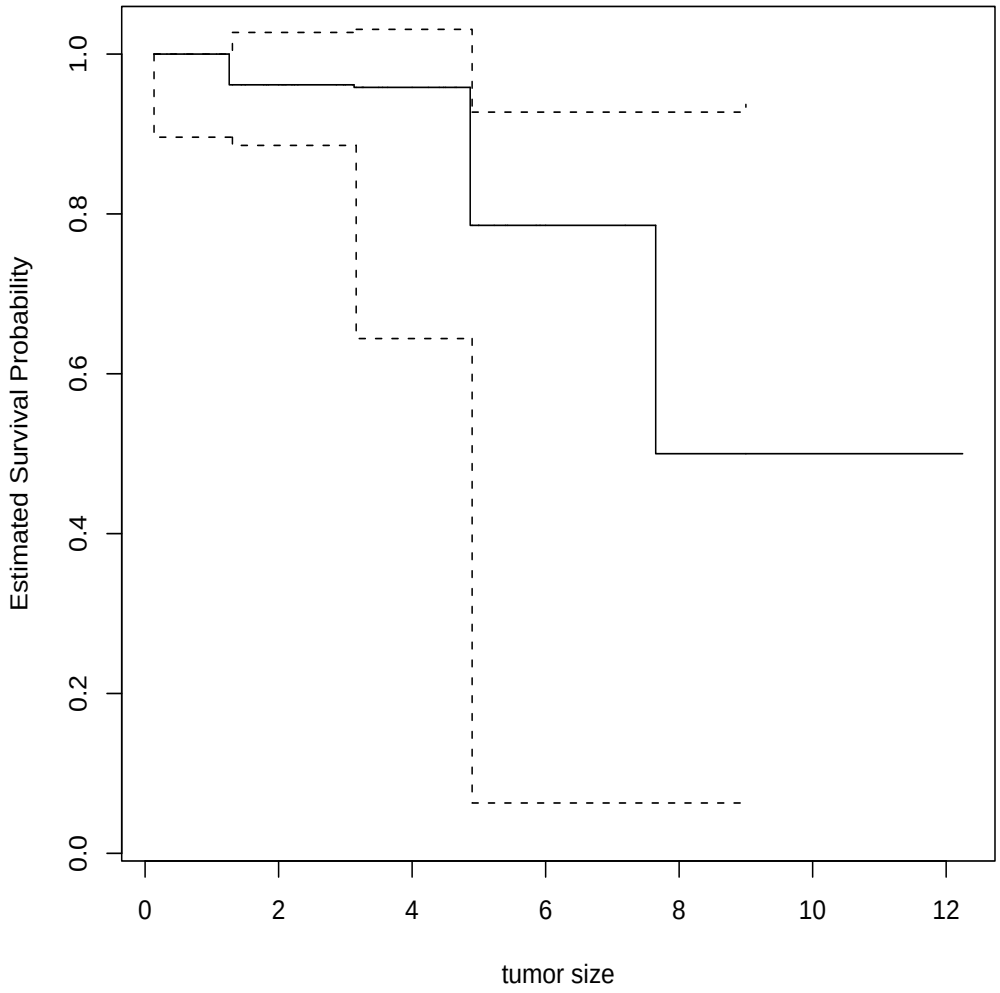

Figure 4: Distribution of survival function for tumor size under Case II scenario for lung epidermoid data (solid line) and 95% pointwise confidence intervals (dashed lines).

Table 1. *Hypothesis testing results for lung screening data*

ρ	p-value
0	0.167
0.25	0.164
0.5	0.163
0.75	0.163
1	0.163

Note: ρ denotes the weight function used in the Harrington & Fleming (1982) procedure.