STATISTICS IN MEDICINE
Statist. Med. (in press)
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2363

Modelling tumour biology–progression relationships in screening trials

Debashis Ghosh*,†

Department of Biostatistics, University of Michigan, Ann Arbor, MI, U.S.A.

SUMMARY

There has been some recent work in the statistical literature for modelling the relationship between tumour biology properties and tumour progression in screening trials. While non-parametric methods have been proposed for estimation of the tumour 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 non-parametric and semi-parametric analysis of modelling tumour size-metastasis data in this article. An association between the models considered by previous authors with survival data structures is discussed. Based on this relationship, we develop non-parametric testing procedures and semi-parametric regression methodology of modelling the effect of size of tumour on the probability at which metastatic transitions occur in two situations. Asymptotic properties of these estimators are provided. The proposed methodology is applied to data from a screening study in lung cancer. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: additive risk model; censoring; interval censoring; monotonicity; non-regular asymptotics; order-restricted inference

BACKGROUND

There has been a rich literature existing on statistical models for tumour progression [1–3] in which the phenotype considered was size of the tumour. Solid cancers develop through a process in which tumours originate as a progenitor cell, which grows to a local lesion that shed cancer cells into the lymphatic system and/or blood stream [4]. 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

*Correspondence to: Debashis Ghosh, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Room M4057, Ann Arbor, MI 48109-2029, U.S.A.
†E-mail: ghoshd@umich.edu

Contract/grant sponsor: National Institutes of Health; contract/grant number: 5P30 CA46592

Received 15 January 2005
Accepted 20 July 2005

with worse clinical prognosis. It is thus of vital scientific interest to understand the relationship between tumour size and probability of detectable metastases. This also has implications for the development of screening programs.

Most of the work in this area has focused on non-parametric estimation of the distribution function of tumour size at which metastatic transitions occur. By metastatic transition, what we are really referring to is the transition to metastasis that would be detectable at diagnosis. As Kimmel and Flehinger [1] write, 'pragmatically, we consider the point at which metastases first become detectable by techniques standardly used in the medical community as the point of metastatic transition.' The data that typically exist in these settings are the sizes of tumour samples and an indicator of presence of detectable metastases. Since the data are collected cross-sectionally, the distribution function of tumour size at which metastatic transitions occur is non-identifiable. Under certain assumptions made by previous authors [1–3], this quantity becomes identifiable. In this work, we focus primarily on the proposal of Kimmel and Flehinger [1], who developed various non-parametric estimation procedures for the distribution function of the tumour size at which detectable 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 tumour or age of the patient can affect the relationship between tumour size and probability of detectable metastases. Knowing if there exists such a difference might lead to different treatment and/or follow-up regimens. It would thus be desirable to have semi-parametric regression models for analysing such data. However, no such models currently exist.

In this article, we develop a comprehensive approach to the analysis of data on tumour size and metastases. More generally, the methods described in this paper can be used to study the relationship between biological properties of tumours and clinical progression in screening studies. A crucial step in our methodology is the demonstration of the relationship between the observed data structures with those from survival data analysis. Based on the formulation, we can use existing results to derive asymptotic results for previously proposed estimators in the literature and formulate hypothesis testing methods and regression generalizations for analysing data on tumour size and metastases that incorporates other covariates. The structure of the paper is as follows. We first consider the model assumptions in Reference [1] about tumour size and progression and relate it to data structures in survival analysis. We then present two scenarios in which the distribution function of tumour size at which detectable metastatic transition occurs is identifiable. Asymptotic results for the one-sample distribution function estimators are then provided. We also develop hypothesis testing procedures and regression models and estimation procedures for the two scenarios. The proposed methodology is illustrated with an example from a lung cancer data screening study. Finally, we conclude with some brief discussion.

NOTATION AND PRELIMINARIES

Data structure and model assumptions

Let V denote the size of the tumour at detection, Z a p-dimensional vector of covariates and δ be an indicator of tumour metastasis (i.e. $\delta = 1$ if metastases are detected when the primary

tumour is diagnosed, $\delta = 0$ 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 the Background, only (V_i, δ_i) $(i = 1, \ldots, n)$ were available. We will now state the model assumptions utilized in Reference [1]:

(a) Primary cancers grow monotonically, and metastases are irreversible.
(b) We will let Y be the random variable for the tumour size at which transition to detectable metastasis occurs. Let the cdf of Y be denoted by F^Y.
(c) Let $\lambda_1(x)$ denote the hazard function for detecting a cancer with metastasis when the tumour size is x. Let $\lambda_0(x)$ denote the hazard function for detecting a cancer with no metastases when the tumour size is x. Assume that $\lambda_1(x) \geqslant \lambda_0(x)$.

This is also the general framework utilized by Xu and Prorok [2, 3]. Based on assumptions (a)–(c), F^Y is in general non-identifiable. 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, non-parametric estimators of F^Y were developed in Reference [1]. However, no asymptotic results regarding these estimators were given.

Equivalences with censored data structures

We now recast the problem in terms of failure time data structures. In survival analysis, the focus is on modelling the distribution to time to event. Let us define T to be the time to detectable metastatic transition (i.e. the time at which the tumour goes from being non-metastatic to metastatic and can be detected by standard diagnostic methods). We will take the starting point of T to be the initiation of the tumour. The assumptions in the previous section allow Y to be treated as a failure time variable. What this implies is that the tumour size will define an alternative time scale relative to that defined by T. The time scale defined by Y will measure progression on a non-linear scale relative to that defined by T. We might be scientifically interested in the time scale defined by Y in and of itself; that is the position taken in this paper.

Note that another key assumption is that we are referring to data collected from a screening trial. Thus, we are testing asymptomatic individuals in a population for presence of a disease. The framework defined in the previous setting might not work in other clinical settings. For example, let us consider prostate cancer. Typically, men receive hormone therapy after being biopsied for prostate cancer; the entire tumour is then resected during surgery (this is referred to as radical prostatectomy). If the tumour sizes of the individual subjects are measured at this time, then assumption that tumour size can be treated as a failure time variable would be questionable because of the fact that the tumours were subject to treatment.

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 $\delta = I(Y \leqslant C)$, 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 [5] 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 [6]. Now the definition of δ is $\delta \equiv I(Y \leqslant V)$. 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 non-parametric estimators of tumour size at which metastatic distributions occur.

STATISTICAL METHODOLOGY

In this section, we describe the appropriate statistical methodology for the case I and II scenarios. As noted before, the tumour size for these situations can be treated as survival times, so existing methodologies for the analysis of survival data can be applied to the tumour size. We now develop non-parametric and semi-parametric procedures for modelling tumour 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 [7]

$$\lambda(y|Z) = \lambda_0(y) \exp(\alpha_0' Z) \tag{1}$$

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 (1), α has a relative risk interpretation on a logarithmic scale, and estimation of α has been well-studied for both right-censored data [7, 8] and for interval-censored data [9]. However, for simplicity of computation, the additive risk model [10, pp. 53–57] is used here

$$\lambda(y|Z) = \lambda_0(y) + \beta_0' Z \tag{2}$$

where β_0 is a p-dimensional vector of unknown regression coefficients. We comment on the use of the proportional hazards model (1) in the discussion.

The second comment deals with the issue of non-measured tumours. In the framework of Reference [1], there was the possibility that there were tumours, both with and without metastases, for which the tumour size was not measured. In the framework presented earlier, this corresponds to observations with missing observed failure time measurements. In this article, we assume that the proportion of non-measured tumours relative to the total number of tumours is asymptotically negligible. This will imply that the limiting distributions of the estimators proposed here with those proposed in Reference [1] will be equivalent. In order to develop estimation procedures with non-measured tumours, some type of imputation method [11] would be required. While Kimmel and Flehinger [1] use an imputation method for incorporating information on non-measured tumours, it requires a missing at random assumption that we believe may not be valid. It seems reasonable to assume that whether a tumour was measured would depend on the size of the tumour, which then becomes a non-ignorable missing data mechanism. How to account for this is beyond the scope of the paper.

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 non-parametric estimation of the survival function corresponding to F^Y, say S^Y, the Kaplan–Meier method can be used. Let $v_{(1)} < v_{(2)} < \cdots < v_{(m)}$ denote the sorted tumour sizes. Then the Kaplan–Meier estimator is given by

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

where d_i is the number of tumours with metastases with size $v_{(i)}$, and n_i is the number of tumours 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 in Reference [1] and is asymptotically equivalent to the estimator of Reference [2] in the case where $C = 1$ (in their notation).

Before proving asymptotic results about the estimator $\hat{S}^Y$, we introduce some notation. Let $N_i(t) = I(V_i \leqslant t, \delta_i = 1)$ and $R_i(t) = I(V_i \geqslant t)$, $i = 1, \ldots, n$. The following result can be proven by standard survival analysis techniques [5]:

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

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

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 $K \geqslant 1$. This corresponds to comparing the distribution function for tumour size at metastatic transition across K groups. The data can be expressed as $\{V_{ij}, \delta_{ij}, i\}$, $j = 1, \ldots, n_i$ and $i = 1, \ldots, K$. We define $N_{ij}(t) = I(V_{ij} \leqslant t, \delta_{ij} = 1)$, $R_{ij}(t) = I(V_{ij} \geqslant t)$, $R_{i\cdot}(t) = \sum_{j=1}^{n_i} R_{ij}(t)$, $N_{i\cdot}(t) = \sum_{j=1}^{n_i} N_{ij}(t)$, $R_{\cdot\cdot}(t) = \sum_{i=1}^{n} R_{i\cdot}(t)$, and $N_{\cdot\cdot}(t) = \sum_{i=1}^{n} N_{i\cdot}(t)$. In order to test $H_0 : F_1^Y = \cdots = F_K^Y$, we can utilize the G^ρ family of test statistics proposed by Harrington and Fleming [12]

$$T = Z'\Sigma Z$$

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

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

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

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

and $\delta_{lm}=I(Z_l=Z_m=l)$. Using arguments in Section 5.2 of Reference [8], 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 semi-parametric model for the effect of covariates on the hazard function through equation (2). Estimation in this model has been previously developed [13]. The following estimating function can be used for estimation of β in (2):

$$U(\beta)=\sum_{i=1}^n\int_0^\infty\{Z_i-\bar{Z}(t)\}\{\mathrm{d}N_i(t)-R_i(t)\beta' Z_i\,\mathrm{d}t\} \tag{3}$$

where $\bar{Z}(t)=\sum_{j=1}^n R_j(t)Z_j/\sum_{j=1}^n R_j(t)$. Setting $U(\beta)$ from (3) equal to zero yields a closed-form estimator for β_0

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

where $a^{\otimes 2}=aa'$. By standard martingale arguments [13], 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

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

and

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

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 can be made [5, pp. 38–40]. Let $v_{(1)}\leqslant v_{(2)}\leqslant\cdots\leqslant v_{(n)}$ 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 non-parametric 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

$$0\leqslant x_1\leqslant\cdots\leqslant x_n\leqslant 1$$

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' [5, p. 40]

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

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

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

Then $\tilde{x}_i$ is the left derivative of H^* at $i=0,\ldots,n$. This estimator corresponds to that proposed in Reference [1] in the case II scenario. They provided no asymptotic analysis of this estimator. Using the arguments in Chapter 5 of Reference [5], we can prove the following result.

Theorem 2
Define $G(t)=\Pr(V\leqslant t)$. 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 $Z\equiv\operatorname{argmin}\{W(t)+t^2\}$, 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 per cent confidence interval for $F^Y(z_0)$ is then given by

$$\{\hat{F}^Y(z_0)-n^{-1/3}\,\hat{Q}_{0.975},\,\hat{F}^Y(z_0)+n^{-1/3}\,\hat{Q}_{0.975}\}$$

where $\hat{Q}_{0.975}$ is a consistent estimator of $Q_{0.975}$, the 97.5th percentile of the limiting random variable CZ. But $Q_{0.975}$ is simply $C\times 0.99818$ where 0.99818 is the 97.5th percentile of Z, where the quantiles of Z are from Reference [14]. Since C involves the unknown parameters $G(z_0)$, $h(z_0)$, and $g(z_0)$, we estimate C by

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

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

$$\left\{\hat{F}^Y(z_0) - n^{-1/3}\,\hat{C}_n \times 0.99818,\ \hat{F}^Y(z_0) + n^{-1/3}\,\hat{C}_n \times 0.99818\right\}$$

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 non-parametric 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 [15].

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 Reference [16]. 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 $W_i(t) \equiv I(Y_i \leqslant t)$. A test of the null hypothesis $H_0 : F_1^Y = \cdots F_K^Y$ is given by the statistic

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

where $\bar{Z} = n^{-1} \sum_{j=1}^{n} Z_j$. It is shown in Reference [16] that under the null hypothesis, $n^{-1/2}\tilde{T}$ 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 (2) in the case II scenario. Lin *et al.* [17] proposed a procedure for estimation of β_0 in model (1) based on the partial likelihood using the counting process $\tilde{N}_{1i}(t) \equiv (1 - \delta_i) I(S_i \leqslant t)$, $i = 1, \ldots, n$. Let $\mathrm{d}H_i(t)$ denote this hazard corresponding to $\mathrm{d}\tilde{N}_{1i}(t)$, the increment in $\tilde{N}_{1i}(t)$, $i = 1, \ldots, n$. Lin *et al.* [17] show that under model (2), the following model for $\mathrm{d}H_i(t)$ $(i = 1, \ldots, n)$ is induced

$$\mathrm{d}H_i(t) = \mathrm{d}H_0(t) \exp\{\beta_0' Z_i^*(t)\} \tag{4}$$

where $\mathrm{d}H_0(t) = \exp\{-\Lambda_0(t)\}\,\mathrm{d}\Lambda^S(t)$, $\Lambda^S(t)$ is the cumulative hazard function of S, $\Lambda_0(t) = \int_0^t \lambda_0(u)\,\mathrm{d}u$, and $Z_i^*(t) = -tZ_i$. The model in (4) has a form identical to that of the proportional hazards model [6]. We can estimate β_0 using the partial likelihood score function

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\tau} \left\{ Z_i^*(t) - \frac{S^{(1)}(\beta, t)}{S^{(0)}(\beta, t)} \right\} \mathrm{d}\tilde{N}_{1i}(t) \tag{5}$$

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 $\hat{\beta}$ be the solution from setting $U(\beta)$ in (5) equal to zero. Using martingale theory, $n^{1/2}(\hat{\beta} - \beta_0)$ converges in distribution to a normal random vector with mean zero and variance $\mathscr{I}(\beta_0) \equiv \lim_{n \to \infty} n^{-1} I(\beta_0)$, where

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

This variance can be consistently estimated based on empirical quantities.

NUMERICAL EXAMPLE

In this section, we consider data from a screening trial involving lung cancer and reported in Reference [1]. 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 raw data are shown in Figure 1.

The proposed techniques are now applied. First, the tumour size and metastases data are analysed under the case I situation, i.e. metastases occur at the time of detection. In this instance, we can treat the tumour size as a right-censored variable. The estimated survival functions for the size distribution and pointwise 95 per cent confidence intervals for the adenocarcinomas and for the epidermoid cancers are given in Figures 2(a) and (c), respectively. Next, differences in the size distribution between the two types of tumours was tested using the Fleming–Harrington G^{ρ} class of statistics. Results for various values of ρ are given in Table I. We find that there is slight evidence of a difference in size distributions between the two types of lung tumours. Finally, we analysed the data using additive risk model (2) in which there is one covariate Z, a binary indicator for tumour 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.

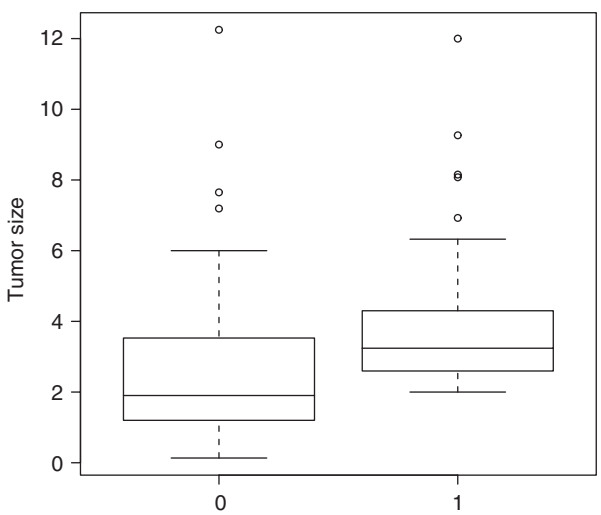

Figure 1. Box plot of tumour sizes for lung cancer screening data. For the horizontal axis, 0 denotes adenocarcinoma and 1 represents epidermoid.

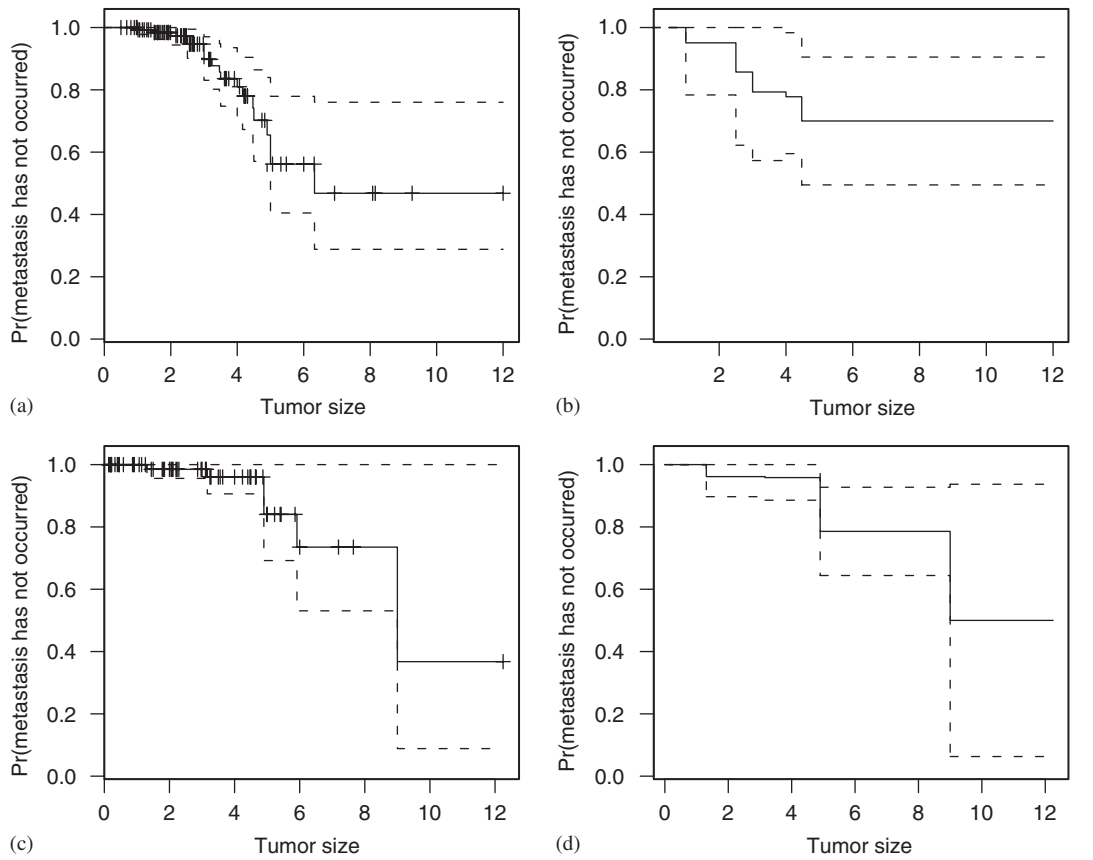

Figure 2. (a) Distribution of survival function for tumour size under case I scenario for lung adenocarcinoma data (solid line) and 95 per cent pointwise confidence intervals (dashed lines). (b) Distribution of survival function for tumour size under case II scenario for lung adenocarcinoma data (solid line) and 95 per cent pointwise confidence intervals (dashed lines). (c) Distribution of survival function for tumour size under case I scenario for lung epidermoid data (solid line) and 95 per cent pointwise confidence intervals (dashed lines). (d) Distribution of survival function for tumour size under case II scenario for lung epidermoid data (solid line) and 95 per cent pointwise confidence intervals (dashed lines).

Next, we analysed the data using the case II scenario; here, the tumour size is now an interval censored random variable. First, we plot the survival functions for the tumour size distributions and associated pointwise 95 per cent confidence intervals; these graphs for the adenocarcinoma and epidermoid lung cancers are given in Figures 2(b) and (d), respectively. In testing for a difference in tumour size distributions between tumour type (adenocarcinoma *versus* epidermoid), the method of Reference [16] yields a test statistic of 1.60 and an associated p-value of 0.11. Finally, the estimation procedure for the semi-parametric additive hazards model of Reference [17] with Z as a binary indicator for tumour site yields an estimated regression coefficient of 0.16 and standard error of 0.10. The Wald statistic gives a

Table I. Hypothesis testing results for lung screening data. Note that ρ denotes the weight function used in the Harrington and Fleming (1982) procedure.

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

corresponding p-value of 0.11, which again suggests a marginal association between tumour type and hazard of tumour size detection. A difference in metastatic distributions between lung cancers from the two sites of origins suggests that there may be different biological mechanisms that dictate the progression to metastasis. This in turn might lead to different treatment regimens for lung adenocarcinomas *versus* lung epidermoids.

We now highlight the differences between our analysis with that in Reference [1]. First, we have ignored the tumours on which there were missing measurements, while Kimmel and Flehinger [1] use a reweighting scheme to impute information from these tumours. We discuss the appropriateness of the reweighting in the Discussion. However, the graphs in Figure 2 are not qualitatively that different from those in Reference [1]. In addition, our theory allows for such standard errors, while no standard errors were derived in Reference [1]. In addition, the testing and regression results presented here are completely new for this problem.

DISCUSSION

In this article, we have laid out a framework for the analysis of data on tumour 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 non-parametric maximum likelihood estimators of the distribution function for tumour 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.

Two cases were considered in the paper. They represent the extremes as to how much information is available in the data about the tumour size at which metastasis detectable at diagnosis occurs. Case I corresponds to the right-censored case in which there is a positive probability of observing the quantity, while case II is the situation where it is never directly observable. Thus in some sense, case I corresponds to the most optimistic analysis, while case II is the least optimistic analysis, where optimality is defined in terms of the amount of the available information. This explains why the standard errors are bigger in case II than in case I. Interestingly, however, the testing and regression procedures gave very similar inferences. One explanation for this result is that even for the case II situation, the standard $n^{1/2}$ asymptotics apply to finite-dimensional parameters such as regression coefficients. This issue is one that needs to be addressed in future work. In terms of plausibility of whether

case I or case II is more plausible, from a point of view of conservatism, we prefer the latter. A more crucial assumption is the existence of Y, the tumour size at which transition to metastasis detectable at diagnosis happens. The methods and theory in this paper are based on that random variable.

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 existing methodologies [7–9].

The major contribution of this paper is to provide a new means to thinking about biological data from screening studies. Provided that one can assume that assumptions similar to those made by Kimmel and Flehinger [1] can hold, we can model biological data as failure time variables, subject to various censoring mechanisms. This would allow the importation of existing survival analysis techniques, as well as potential generalizations. For example, if the biological response variable modelled was bivariate, then one might entertain multivariate survival analysis techniques for analysing the data. However, a key point is the monotonicity described earlier; this allows for the equivalence with survival data structures. If subjects are given treatment, then the assumption of treating the tumour size as a failure time variable becomes suspect.

As mentioned before, in many cancer screening studies, some tumours are not measured. A next logical question is whether one can incorporate information from these tumours into estimation of the distribution function for tumour size when metastasis can be diagnosed. In Reference [1], they are incorporated using a reweighting scheme that is valid when the missingness of tumour size occurs at random. However, the missing at random seems implausible, as one might expect sizes of smaller tumours to be more difficult to measure. How to adjust for this is an area we are currently pursuing.

Much of the literature in the area of cancer screening has focused on mechanistic models for tumour progression [18]. Such models tend to be parametric in nature and potentially have identifiability issues, while the methods we have proposed here are less parametric. These procedures proposed in the paper could serve as an exploratory device in order to determine what parametric models can be utilized.

ACKNOWLEDGEMENTS

The author would like to thank Dr Marek Kimmel for providing the lung cancer data and two referees, whose comments greatly improved the manuscript. This research is supported in part by the National Institutes of Health through the University of Michigan's Cancer Center Support Grant (5P30 CA46592).

REFERENCES

1. Kimmel M, Flehinger BJ. Nonparametric estimation of the size-metastasis relationship in solid cancers. *Biometrics* 1991; **47**:987–1004.
2. Xu JL, Prorok PC. Nonparametric estimation of solid cancer size at metastasis and probability of presenting with metastasis at detection. *Biometrics* 1997; **53**:579–591.
3. Xu JL, Prorok PC. Estimating a distribution function of the tumour size at metastasis. *Biometrics* 1998; **54**:859–864.
4. Foulds L. Characteristics of neoplasms. In *Neoplastic Development*, Foulds L (ed.). Academic Press: London, 1969; 97–136.
5. Fleming T, Harrington DP. *Counting Processes and Survival Analysis*. Wiley: New York, 1991.
6. Groenenboom P, Wellner JA. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser: Basel, 1992.

7. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society*, *Series B* 1972; **34**:187–220.
8. Andersen PK, Borgan O, Gill RD, Keiding N. *Statistical Models Based on Counting Processes*. Springer: New York, 1993.
9. Huang J. Efficient estimation for the proportional hazards model with interval censoring. *Annals of Statistics* 1996; **24**:540–568.
10. Breslow NE, Day NE. *Statistical Methods in Cancer Research* (*Vol. 1*): *The Analysis of Case–control Studies*. Lyon: IARC, 1980.
11. Little RJA, Rubin DB. *Statistical Analysis with Missing Data* (2nd edn). Wiley: New York, 2002.
12. Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**:133–143.
13. Lin DY, Ying Z. Semiparametric analysis of the additive risk model. *Biometrika* 1994; **81**:61–71.
14. Groeneboom P, Wellner JA. Computing Chernoff's distribution. *Journal of Computational and Graphical Statistics* 2001; **10**:388–400.
15. Heckman NE, Ramsay JO. Penalized regression with model-based penalties. *Canadian Journal of Statistics* 2000; **28**:241–258.
16. Sun J. A nonparametric test for current status data with unequal censoring. *Journal of the Royal Statistical Society*, *Series B* 2000; **61**:243–250.
17. Lin DY, Oakes D, Ying Z. Additive hazards regression with current status data. *Biometrika* 1998; **85**:289–298.
18. Yakovlev AY, Tsodikov AD. *Stochastic Models of Tumor Latency and their Biostatistical Applications*. World Scientific Press: Singapore, 1996.

