Biometrics 64, 141–148
March 2008

DOI: 10.1111/j.1541-0420.2007.00830.x

Proportional Hazards Regression for Cancer Studies

Debashis Ghosh

Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, Michigan 48109-2029, U.S.A.
email: ghoshd@umich.edu

Summary. There has been some recent work in the statistical literature for modeling the relationship between the size of cancers and probability of detecting metastasis, i.e., aggressive disease. Methods for assessing covariate effects in these studies are limited. In this article, we formulate the problem as assessing covariate effects on a right-censored variable subject to two types of sampling bias. The first is the length-biased sampling that is inherent in screening studies; the second is the two-phase design in which a fraction of tumors are measured. We construct estimation procedures for the proportional hazards model that account for these two sampling issues. In addition, a Nelson–Aalen type estimator is proposed as a summary statistic. Asymptotic results for the regression methodology are provided. The methods are illustrated by application to data from an observational cancer study as well as to simulated data.

Key words: Biased sampling; Empirical process; Inverse probability weighting; Natural history; Screening program.

1. Introduction

Given the morbidity and mortality and associated costs of treating people with cancer, it is important to determine optimal screening schedules for early detection of cancer. There has been much work done on developing mathematical models of screening (Yakovlev and Tsodikov, 1996, Chapter 5). Much of the previous work in this area has focused on consideration of multistate models (Zelen and Feinleib, 1969; Albert, Gertman, and Louis, 1978; Day and Walter, 1984; Shen and Zelen, 1999). Based on the estimated parameters from the model, one can then begin to assess effects of screening interventions on these quantities. More recent work has focused on estimating operating characteristics of screening programs (e.g., Baker, Erwin, and Kramer, 2003).

An alternative approach is to better understand the relationship between various aspects of tumor biology and progression. Such procedures have been proposed by Kimmel and Flehinger (1991) and Xu and Prorok (1997, 1998) in the one-sample setting. In many situations, it is of interest to determine the effects of covariates on tumor progression. Such a question requires a regression formulation. Recently, an approach to this problem has been outlined by Ghosh (2006). The regression model utilized there was the additive hazards model (Breslow and Day, 1980; pp. 53–57). However, much more popular for the analysis of failure time data is the proportional hazards regression model (Cox, 1972). For that model, the regression parameters are interpretable as relative risk ratios on a logarithmic scale, which are easily comprehended by epidemiologists.

One issue with tumor data from screening trials is that of length-biased sampling (Zelen and Feinleib, 1969; Patil and Rao, 1978). Because of lead-time bias due to screening, the types of tumors that are detected tend to be smaller and slower-growing tumors. In the initial work of Kimmel and Flehinger (1991), it was assumed that there was no biased sampling. While the proportional hazards regression model for length-biased data has been studied by Wang (1996), she did not consider the presence of censoring, which is present in the data we consider here.

Another complication in the collection of tumor data is that only a fraction of tumors have size measurements. The way this was incorporated by Kimmel and Flehinger (1991) is to assume that the missingness mechanism for the distribution of tumor sizes is missing completely at random (Little and Rubin, 2002). This issue is not addressed in the statistical framework of Ghosh (2006). It seems that this might not be a completely reasonable assumption. A more realistic scenario is to assume that the tumors that tend to be measured are a function of observed covariates.

In this article, we develop estimation procedures for the analysis of tumor size–metastasis data in cancer studies. The course of this article is as follows. In Section 2, we state the model for tumor size and progression and briefly review the results of Kimmel and Flehinger (1991). We will also discuss the notion of time scales and induced dependent censoring for this problem. It turns out that some further assumptions are needed to ensure the validity of one class of the estimators proposed by Kimmel and Flehinger (1991). Semiparametric inference procedures for the proportional hazards model with length-biased data under a deterministic growth assumption will be outlined in Section 3. While the method of Wang (1996) has been proposed for length-biased data, it cannot be directly applied for censored data. In this article, we construct a new estimation procedure based on estimating equations; asymptotic results are provided as well. We also describe the extension to accommodate nonmeasured tumors.

The proposed methods are then applied to data from a lung cancer screening study in Section 4. The finite-sample procedures of the proposed methods are also assessed using a small simulation study. In addition, we assess the sensitivity of inferences to violations in model assumptions using simulation studies as well. Finally, we conclude with some discussion in Section 5.

2. Tumor Screening Framework

2.1 *Notation and Model Assumptions*

Let S denote the tumor size, δ be an indicator of metastasis (i.e., $\delta = 1$ if the tumor has metastasis and zero otherwise) and $\mathbf{Z}$ represent a $p \times 1$ vector of covariates. For $i = 1, \ldots, n$, the observed data consist of $\{S_i,\ \delta_i,\ \mathbf{Z}_i\}$, n independent copies from $\{S,\ \delta,\ \mathbf{Z}\}$. We will also assume that there is a single screening exam given.

In their framework, Kimmel and Flehinger (1991) make the following model assumptions:

(a) Primary cancers grow monotonically, and metastases are irreversible.
(b) Denote Y as the random variable for the distribution of the primary tumor sizes at which metastatic transitions take place. Let the CDF of Y be denoted by F^Y.
(c) Let $\lambda_1^d(x)$ denote the hazard function for detecting a cancer with metastasis when the tumor size is x. Let $\lambda_0^d(x)$ denote the hazard function for detecting a cancer with no metastases when the tumor size is x. Assume that $\lambda_1^d(x) \geq \lambda_0^d(x)$.

Note that in Assumption (b), we define metastatic transition as the point at which the metastasis can become detectable using existing diagnostic methodologies. Under Assumptions (a)–(c), Y defines a time scale; however, its distribution is not identifiable with the observed data (S, δ, $\mathbf{Z}$). Kimmel and Flehinger (1991) describe two scenarios in which F_Y, the distribution function of Y, is identified from the observed data. The first is when cancers are detected immediately when the metastatic transition occurs. The second is when the detection of the cancer is not affected by the presence of metastases. We will refer to these as Models I and II. In Ghosh (2006), it is shown that Model I corresponds to treating S as a right-censored version of Y. Here and in the sequel, only Model I is considered. Conceptually, the data are cross-sectional; we do not consider any followup on the subjects as in other prevalent cohort setting (e.g., Asgharian and Wolfson, 2005 and the references therein).

The effect of $\mathbf{Z}$ on Y is formulated through the proportional hazards model:

$$\lambda(y \,|\, \mathbf{Z}) = \lambda_0(y) \exp\left(\beta_0^T \mathbf{Z}\right), \tag{1}$$

where $\lambda(\cdot \,|\, \mathbf{Z})$ is the hazard function conditional on covariates, $\lambda_0(\cdot)$ is an unspecified baseline hazard function, and $\boldsymbol{\beta}_0$ is a $p \times 1$ vector of unknown regression coefficients.

2.2 *Time Scales and Induced Dependent Censoring*

What the Kimmel–Flehinger framework corresponds to is using a different time scale for analysis. The scale is that defined by the size of the tumor. We briefly discuss the utility of the time scale being used here and bring up a crucial point involving induced censoring.

Let T denote the time at which the tumor becomes metastatic and is potentially detectable by diagnosis. This occurs on the chronological time scale. However, in screening trials, there is a time zero representing start of study, and screening exams occur after the study begins. Thus, the screening exam times are assessed on the study time scale. The assumptions of Kimmel and Flehinger (1991) define $Y \equiv \phi(T)$ as a new time scale. A question then becomes in what sense will Y be fully informative about T. This can be answered by utilizing the notion of an ideal time scale (Duchesne and Lawless, 2000). Y is an ideal time scale if $\Pr(Y > y) = \Pr(T > \phi^{-1}(y))$ and ϕ is a monotonic function, i.e., T and Y generate the same survivor functions.

It turns out that the validity of the estimation procedures proposed by Kimmel and Flehinger rests on the following additional assumptions:

A. ϕ is a deterministic and monotonic function of time.
B. There exists a single exam time.
C. The distribution of time of progression to metastasis is independent of age at study entry.

Assumption A is needed to avoid the problem of induced dependent censoring. Intuitively, tumor size is really a stochastic process over time. If C denotes time to the screening exam, then what we observe is $S(C \wedge T)$, which might still be dependent even if T and C are independent. Condition A is sufficient to guarantee independence and implies that Y and T are ideal time scales as described above. This implies that violation of Assumption A leads to bias in the Kimmel–Flehinger estimator of the distribution tumor size at metastasis if no information on time is provided. However, this assumption is restrictive; it means that if two tumors have the same time to metastasis, then they have the same tumor size. This seems to be a highly implausible assumption; in the Discussion, we discuss relaxation of this assumption.

The second condition is needed to provide a simple definition of the censoring time here. Potentially, this assumption can be relaxed if screening is defined by a schedule of examination times $t_1, \ldots, t_M$. Then, we would need to assume that the examination time process is independent of T.

Finally, Condition C is needed to account for the fact that subjects who enter the screening trial may have tumors that are in the undetectable preclinical phase. In fact, the estimators proposed by Kimmel and Flehinger (1991) presume that the induced time scale for size equals zero at the start of the study. This is more restrictive than Condition C, a more realistic assumption that would not affect the validity of the Kimmel–Flehinger estimators. In this instance, the Kimmel–Flehinger estimators are in effect estimating a truncated distribution function of tumor size at the metastatic transition point. A violation of the assumption is when the distribution of tumor sizes is age dependent. This again renders the estimation methodology invalid.

2.3 *Length-Biased Sampling*

Another complication in analyzing tumor size data is the presence of length-biased sampling. What this means is that the tumors detected in a screening program tend to be slower-growing tumors. In terms of the probability model being considered, we now have the additional sampling mechanism to deal with, in addition to (1):

$$f_l(y \mid \mathbf{Z}) = \frac{w(y) f(y \mid \mathbf{Z})}{\int w(y) f(y \mid \mathbf{Z})\, dy}, \tag{2}$$

where $w(y)$ is a known positive and increasing weight function. In the absence of covariates, one-sample estimation procedures have been considered by Cox (1969), Vardi (1982), and Jones (1991). The semiparametric model (1) subject to length-biased sampling, with $w(y) = y$ in (2), has been considered by Wang (1996). However, in that proposal, there was no censoring. In addition, we have to deal with the issue that only a fraction of tumors are measured.

The probabilistic model (2) with $w(y) = y$ is consistent with the distribution of interevent times in the so-called positive stable disease model (Day and Walter, 1984). We study the effects of misspecification in the simulation studies in Section 5.

3. Estimation Procedures for Uncensored and Censored Data

3.1 *Regression Methodology*

In this section, we will describe estimation procedures for the semiparametric proportional hazards regression model assuming that tumor size is a monotonic and deterministic function of time to tumor metastasis. We first start with the case of all tumors being measured, followed by the situation where a subset of tumors is measured. While the problem of estimation in (1) with length-biased data has been considered by Wang (1996), there was no censoring in her situation (i.e., $Y \leq C$ with probability one). In the data we collect, an added complication is that censoring is present. The censoring represents tumors detected by screening but which do not have metastases. As will be shown later, the estimation procedure of Wang (1996) cannot be directly applied for this situation.

To construct the proposed estimation procedure, we begin by considering the case of no censoring. We slightly generalize the results of Wang (1996) to an increasing and positive function $w(y)$. Let $N_i^*(t) = I(Y_i \leq t)$ and $R_i^*(t) = I(Y_i \geq t)$, $i = 1, \ldots, n$. In Wang (1996), it is shown that the process

$$M_i^*(t;\beta) = N_i^*(t) - \int_0^t \frac{u}{w(Y_i)} R_i^*(u) \exp\left(\beta^T \mathbf{Z}_i\right) \lambda_0(u)\, du \quad (i = 1, \ldots, n) \tag{3}$$

is a martingale. The quantity (3) motivates the following estimating function for estimation of β in (1), proposed by Wang (1996):

$$\mathbf{U}(\beta) = \sum_{i=1}^{n} \int_0^\infty \left\{ \mathbf{Z}_i - \frac{\sum_{j=1}^{n} t\{w(Y_j)\}^{-1} R_j^*(t) \mathbf{Z}_j \exp\left(\beta^T \mathbf{Z}_j\right)}{\sum_{j=1}^{n} t\{w(Y_j)\}^{-1} R_j^*(t) \exp\left(\beta^T \mathbf{Z}_j\right)} \right\} dN_i^*(t). \tag{4}$$

Define $\hat{\beta}$ to be the estimator derived from setting (4) equal to zero. By standard martingale arguments (Andersen et al., 1993), 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 $\mathbf{A}^{-1}$, where $\mathbf{A}$ is consistently estimated by

$$\hat{\mathbf{A}} = n^{-1} \sum_{i=1}^{n} \times \int_0^\infty \left\{ \mathbf{Z}_i - \frac{\sum_{j=1}^{n} t\{w(Y_j)\}^{-1} R_j^*(t) \mathbf{Z}_j \exp(\beta^T \mathbf{Z}_j)}{\sum_{j=1}^{n} t\{w(Y_j)\}^{-1} R_j^*(t) \exp(\beta^T \mathbf{Z}_j)} \right\}^{\otimes 2} dN_i^*(t),$$

with $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$. We now consider the situation where both Y and C are known. Note that this is an artificial situation, but it will motivate our proposed estimation procedure. For this scenario, the process $M_i(t;\beta_0) \equiv \int_0^t I(C_i \geq u)\, dM_i(t;\beta_0)$ $(i = 1, \ldots, n)$ will also be a zero-mean process. Simple algebraic manipulations, mimicking those in Wang (1996), would reveal

$$\mathbf{V}(\beta) \equiv \sum_{i=1}^{n} \int_0^\infty \left\{ \mathbf{Z}_i - \frac{\sum_{j=1}^{n} I(S_j \geq t) t\{w(Y_j)\}^{-1} \mathbf{Z}_j \exp\left(\beta^T \mathbf{Z}_j\right)}{\sum_{j=1}^{n} I(S_j \geq t) t\{w(Y_j)\}^{-1} \exp\left(\beta^T \mathbf{Z}_j\right)} \right\} dN_i(t) \tag{5}$$

where $N_i(t) = I(Y_i \leq t, \delta_i = 1)$, to be a valid estimating function for β. However, because we do not observe either Y or C completely, (5) cannot be used as an estimating function for β based on the observed data. To be specific, if we replace $w(Y)$ by $w(S)$ in (5), then it is no longer a zero-mean estimating function for β. Our approach is to modify it; we need to replace $I(S_i \geq t)/w(Y_i)$ in (5) by an observable quantity with the same expectation. Note the following two facts: first, that

$$\begin{aligned} E\left\{ \frac{I(S_i \geq t)}{w(Y_i)} \right\} &= E\left[E\left\{ \frac{I(S_i \geq t)}{w(Y_i)} \,\middle|\, Y_i, \mathbf{Z}_i \right\} \right] \\ &= E\left[E\left\{ I(C_i \geq t) \frac{I(Y_i \geq t)}{w(Y_i)} \,\middle|\, Y_i, \mathbf{Z}_i \right\} \right] \\ &= G_C(t) E\left\{ \frac{I(Y_i \geq t)}{w(Y_i)} \right\} \end{aligned}$$

second that

$$
\begin{aligned}
&E\left\{\frac{I(S_i \geq t)\delta_i}{w(Y_i)G_C(Y_i)}\right\} \\
&= E\left[E\left\{\frac{I(S_i \geq t)\delta_i}{w(Y_i)G_C(Y_i)}\,\middle|\, Y_i, \mathbf{Z}_i\right\}\right] \\
&= E\left[G_C(Y_i)E\left\{I(C_i \geq t)\frac{I(Y_i \geq t)}{w(Y_i)G_C(Y_i)}\,\middle|\, Y_i, \mathbf{Z}_i\right\}\right] \\
&= G_C(t)E\left\{\frac{I(Y_i \geq t)}{w(Y_i)}\right\},
\end{aligned}
$$

where $G_C(t) \equiv \Pr(C > t)$ is the survival function for C. Thus, the observable quantity $I(S_i \geq t)\delta_i/w(Y_i)G_C(Y_i)$ $(i = 1, \ldots, n)$ has the same expectation as $I(S_i \geq t)/w(Y_i)$. Because G_C is unknown, we replace it by its estimator based on the Kaplan–Meier method. This leads to the following estimating function:

$$
\tilde{\mathbf{U}}(\beta) = \sum_{i=1}^{n}\int_0^{\tau}\{\mathbf{Z}_i - \tilde{\mathbf{Z}}(t;\beta)\}\, dN_i(t), \tag{6}
$$

where

$$
\tilde{\mathbf{Z}}(t;\beta) = \frac{\sum_{j=1}^{n} t\delta_j\{\hat{G}_C(Y_j)w(Y_j)\}^{-1}R_j(t)\mathbf{Z}_j \exp\left(\beta^T\mathbf{Z}_j\right)}{\sum_{j=1}^{n} t\delta_j\{\hat{G}_C(Y_j)w(Y_j)\}^{-1}R_j(t)\exp\left(\beta^T\mathbf{Z}_j\right)},
$$

where $R_j(t) = I(S_j \geq t)$, $j = 1, \ldots, n$. The constant $\tau > 0$ is chosen to satisfy $G_C(\tau) > 0$. Let $\tilde{\beta}$ denote the solution from setting $\tilde{\mathbf{U}}(\beta) = 0$. The relevant zero-mean process corresponding to (6) is

$$
\tilde{M}_i(t;\beta) = N_i(t) - \int_0^t \frac{u\delta_i}{G_C(Y_i)w(Y_i)} I(S_i \geq u)e^{\beta^T\mathbf{Z}_i}\lambda_0(u)\, du, \tag{7}
$$

$i = 1, \ldots, n$. Note that we cannot directly apply martingale theory as before to derive the consistency and asymptotic normality of $\tilde{\beta}$ because of the G_C and δ in (7). However, (7) remains a valid zero-mean process for $\beta = \beta_0$. In the Appendix, given in the Supplementary Materials, we prove the following theorem:

THEOREM 1: *Assuming certain regularity conditions, given in the Appendix, the estimator $\tilde{\beta}$ is strongly consistent, i.e., $\tilde{\beta} \to_{a.s.} \beta_0$. Furthermore, the random vector $n^{1/2}(\tilde{\beta} - \beta_0)$ converges in distribution to a zero-mean normal random vector with a covariance matrix that can be consistently estimated by $\boldsymbol{I}^{-1}\boldsymbol{S}\boldsymbol{I}^{-1}$, where*

$$
\mathbf{I} = -n^{-1}\frac{d\tilde{\mathbf{U}}(\beta)}{d\beta}\,\Bigg|_{\beta=\tilde{\beta}}
$$

and

$$
\mathbf{S} = n^{-1}\sum_{i=1}^{n}\int_0^{\tau}\{\mathbf{Z}_i - \tilde{\mathbf{Z}}(t;\tilde{\beta})\}^{\otimes 2}\, dM_i(t;\tilde{\beta}).
$$

Note that (6) falls into the class of inverse probability of censoring weighted (IPCW) estimating equations studied by Robins and colleagues (e.g., Robins and Rotnitzky, 1992). There are several ways in which (6) could be extended. For example, we can model the censoring based on covariates. For example, if $\mathbf{Z}$ consists of discrete covariates, then we could plug in covariate-specific Kaplan–Meier estimators for censoring into (6). A generalization that adjusts for discrete and continuous covariates is via a stratified proportional hazards model for censoring:

$$
\lambda^C(t \mid \mathbf{Z}, \mathbf{W}) = \lambda_{0w}^C(t)\exp(\gamma^T\mathbf{Z}),
$$

where $\mathbf{W}$ is a set of discrete covariates, the combination of whose values defines the strata for the model. Based on the estimates from the model, one can compute individual-specific censoring probabilities and plug them into (6); this is the approach taken in Robins and Rotnitzky (1992). One could also derive the so-called "doubly robust" estimating functions using techniques from van der Laan and Robins (2003). These extensions are beyond the scope of this article.

A generalization of the estimating equation is to consider the following class of weighted estimating equations:

$$
\tilde{\mathbf{U}}^h(\beta) = \sum_{i=1}^{n}\int_0^{\tau} H(\beta, t)\{\mathbf{Z}_i - \tilde{\mathbf{Z}}(t;\beta)\}\, dN_i(t), \tag{8}
$$

where $H(\beta,\ t)$ is a data-dependent weight function. Ideally, $H(\beta,\ t)$ would be chosen in a manner to maximize the efficiency of the resulting solution from setting (8) equal to zero. However, the optimal weight function in (8) will be difficult to find because it will depend on the weight function in (2) as well as the censoring distribution. Both of these quantities are unknown.

3.2 *Baseline Hazard Estimation*

In usual survival analysis problems with right-censored data, graphical summaries are given by Kaplan–Meier or Nelson–Aalen estimators. For the setting proposed here, such techniques have not been available. We can use (7) to motivate a Nelson–Aalen type estimator that can be used for giving a descriptive statistic. Assume that $\beta_0 = 0$ and G_C is known. Setting $\sum_{i=1}^{n}\int_0^t d\tilde{M}_i(u;0) = 0$ and rearranging terms yield the following solution:

$$
\Lambda_0(t) = \sum_{i=1}^{n}\int_0^t \frac{dN_i(u)}{\sum_{j=1}^{n} u\delta_j I(S_j \geq u)\{w(Y_j)G_C(Y_j)\}^{-1}}. \tag{9}
$$

Plugging in an estimator of G_C into (9) gives a Nelson Aalen-type estimator for $\Lambda_0(t)$. The resulting estimator has a form very similar to that of the usual Nelson–Aalen estimator for right-censored data. The major difference is that the risk set in the denominator of (9) has a much more complicated form to account for the length-biased sampling and censoring.

3.3 *Estimation for Two-Stage Design*

We now incorporate the fact that $(S, \mathbf{Z})$ is not collected on all tumors. The design of the study is treated as two-stage sampling; this type of design is very common in survey sampling studies. Let D denote the indicator of selection of the tumor into the second stage of the study. We make two assumptions about selection into the second stage. First, we assume that D is conditionally independent of S given $\mathbf{Z}$. However, we do

allow the probability of inclusion into the second stage to depend on $\mathbf{Z}$; denote this as $\pi(\mathbf{Z}) \equiv P(D = 1 \mid \mathbf{Z})$. The following estimating function is used for estimation of β in this setting:

$$\tilde{\mathbf{U}}_c(\beta) = \sum_{i=1}^{n} \int_0^{\infty} \pi(\mathbf{Z}_i)\{\mathbf{Z}_i - \tilde{\mathbf{Z}}_{cc}(t;\beta)\}\, dN_i(t),$$

where

$$\tilde{\mathbf{Z}}_{cc}(t;\beta) = \frac{\sum_{j=1}^{n} \pi(\mathbf{Z}_j) t\delta_j \{\hat{G}_C(Y_j) w(Y_j)\}^{-1} R_j(t) \mathbf{Z}_j \exp\left(\beta^T \mathbf{Z}_j\right)}{\sum_{j=1}^{n} \pi(\mathbf{Z}_j) t\delta_j \{\hat{G}_C(Y_j) w(Y_j)\}^{-1} R_j(t) \exp\left(\beta^T \mathbf{Z}_j\right)}$$

for $j = 1, \ldots, n$. Let $\tilde{\beta}_c$ denote the solution from setting $\tilde{\mathbf{U}}_c(\beta) = 0$. It can be shown using arguments in Lin (2000) that $\tilde{\beta}_c$ is consistent for β_0 and that $n^{1/2}(\tilde{\beta}_c - \beta_0)$ is asymptotically normal. The form of the variance of the limiting distribution is similar to that for estimators of the proportional hazards model under case–cohort sampling (Barlow, 1994). We have utilized the jackknife in order to estimate the variance of $\tilde{\beta}_c$ (Lipsitz, Dear, and Zhao, 1994).

4. Numerical Examples

4.1 *Lung Cancer Data*

In this section, we apply the proposed methodology to data from a screening trial involving lung cancer and reported in Kimmel and Flehinger (1991). The lung cancer data were 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. The 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, six have metastases. There are also nonmeasured tumors which account for the two-phase design we described; the missing data measurements are summarized in Table 1. It is of interest to determine if there is an association between site of origin and aggressiveness of tumor. Thus, there is a single covariate Z, site of origin. We code it 0/1 for adenocarcinomas and epidermoids, respectively. The data on tumor sizes are shown in Figure 1. Recall that the tumor size will be used as a failure time variable. Based on the plot, we find evidence that the epidermoids have larger tumor sizes. However, this graph does not incorporate the censoring or the length-biased sampling. In all analyses reported in this section, the weight function $w(y) = y$ is used.

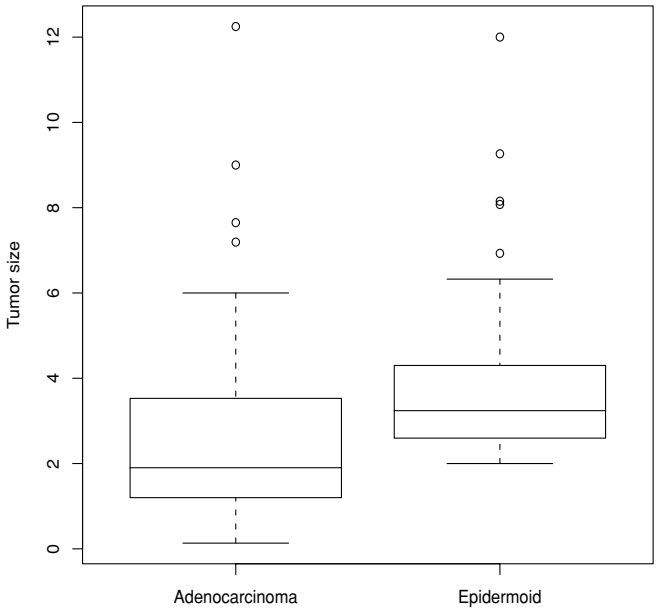

Figure 1. Boxplot of tumor sizes for lung cancer screening data.

Next, we consider Nelson–Aalen estimators of the tumor size distributions; these are given in Figure 2. The plots suggest that epidermoids are at increased risk of metastasizing relative to adenocarcinomas. However, the difference appears less in Figure 2a than in Figure 2b. However, the estimated risk of metastasis for adenocarcinomas in Figure 2b appears to be very close to zero, which suggests that there might be substantial variability in this estimate. Note also that in Figure 2b, there appears to be less evidence for proportional hazards, accounting for the length-biased sampling.

We first use the analysis method outlined in Ghosh (2006). This corresponds to treating the tumor size as a right-censored random variable and fitting a proportional hazards analysis that ignores the biased sampling and the two-stage design. This can be fit using existing survival analysis software for the Cox model and yields an estimated relative risk of $\exp(0.56) = 1.76$ with an associated 95% confidence interval

Table 1
Summary of lung cancer data

	Adenocarcinomas		Epidermoids	
Status	Metastatic	Nonmetastatic	Metastatic	Nonmetastatic
Measured	19	122	6	81
Not measured	15	8	12	12

Note: Status refers to whether or not the tumor size is measured; not measured tumors are treated as missing in the analysis. Cell entries are number of samples under each classification.

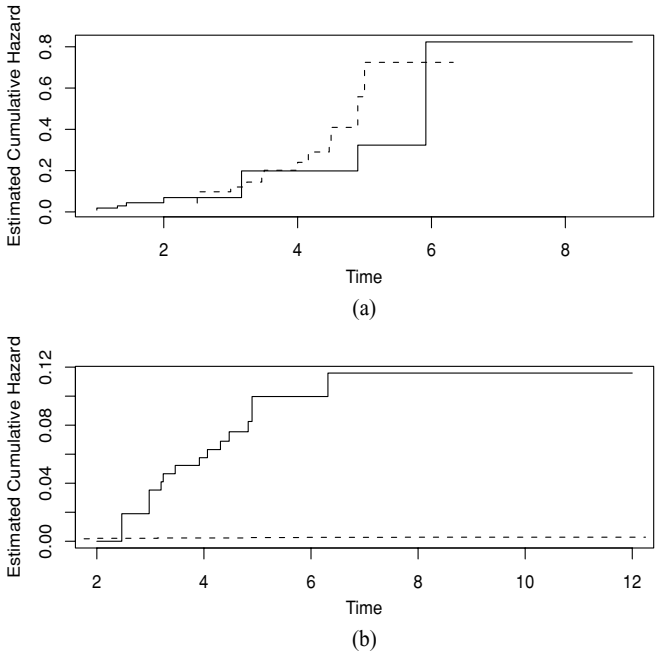

Figure 2. Nelson–Aalen estimators for the cumulative hazard function corresponding to distribution of tumor size at metastatic transition. Solid line is for the epidermoid cancers, while dashed line is for the adenocarcinomas. (a) The estimators when tumor size is treated as right-censored random variable, and the length–biased sampling mechanism is ignored. (b) Based on treating the tumor size as a right-censored random variable with the length-biased sampling mechanism incorporated. All estimators ignore the two-phase sampling.

(CI) of (0.79, 3.93). Thus, while the epidermoid tumors tend to have smaller sizes for detectable metastasis, the association is not statistically significant. Next, we present the analysis which ignores the nonmeasured tumors and incorporates the biased sampling aspect. This corresponds to the method described in Section 3.1. This analysis gives an estimated relative risk of exp(0.16) = 1.17 with an associated 95% CI of (0.51, 2.73). Thus, the epidermoids have increased risk of smaller tumor size for detectable metastasis in this analysis as well. However, incorporating the length-biased sampling mechanism leads to a 33% reduction in the relative risk, although the association is still nonsignificant.

Finally, we present results of regression models in which the two-phase sampling is taken into account. The first analysis ignores site of origin; this corresponds to treating $\pi(Z) = 0.82$. This analysis gives an estimated relative risk of exp(0.12) = 1.13 with an associated 95% CI of (0.52, 2.42). The second analysis allows for different sampling weights based on site of origin; based on the data in Table 3, this yields $\pi(Z) = 0.86$ and 0.78 for adenocarcinomas and epidermoids, respectively. This analysis gives an estimated relative risk of exp(0.15) = 1.16 with an associated 95% CI of (0.55, 2.45). In this instance, it does not appear that incorporating the case–control sampling into the analysis appears to change the answer substantially.

Table 2
Summary of simulation results for β

	U(0,2) censoring				U(0,6) censoring			
n	Bias	SE	SEE	CP	Bias	SE	SEE	CP
50	−0.01	0.59	0.55	0.92	−0.04	0.49	0.45	0.93
100	−0.02	0.39	0.37	0.94	−0.02	0.34	0.32	0.94
200	0.01	0.31	0.30	0.95	−0.01	0.24	0.24	0.95

Note: Bias is the mean of the estimators of θ_0 minus θ_0; SE is the standard error of the estimators of θ_0; SEE is the mean of the standard error estimate; and CP is the coverage probability of the 95% confidence interval.

4.2 *Simulation Studies*

To assess the finite-sample properties of the regression estimation methodology from Section 3.1, we performed a series of simulation studies. We let $\mathbf{Z} \equiv Z$ denote a binary indicator and used the same simulation setup as in Wang (1996). In particular, we took the density for $Z = 0$ to be

$$f_0(y) = 2ye^{-y^2}, \quad y > 0$$

and that for $Z = 1$ to be

$$f_1(y) = 2e^{\beta} ye^{-e^{\beta}y^2}, \quad y > 0.$$

Thus, the hazard ratio between the two groups is β; in the simulations, we considered $\beta = 2$. Length-biased sampling was incorporated using $w(y) = y$ and was independent of the group. Two types of censoring were considered. In the first, $U(0, 2)$ random variables were used for independent censoring, while in the second situation, $U(0, 6)$ random variables were generated. This yielded approximately 50% and 20% censoring, respectively. For each simulation scenario, 500 data sets were generated. Sample sizes $n = 50$, 100, and 200 were considered. The results are given in Table 2. Based on the simulation results, the proposed estimation procedure appears to have satisfactory performance for the sample sizes considered, with diminishing bias for larger sample sizes.

We also studied a situation in which the data are generated using the same mechanism, but in which we fit the estimating equation using $w(y) = y^2$. This was done in order to study the effects of misspecification in the estimation procedure. The same simulation settings were used as before; the results are given in Table 3. Based on the table, we find that for larger sample sizes, the estimators show greater bias and worse coverage probabilities for the 95% confidence intervals.

Table 3
Summary of simulation results for β under misspecification

		U(0,2) censoring			U(0,6) censoring		
$w(y)$	n	Bias	SEE	CP	Bias	SEE	CP
y^2	50	−0.18	0.82	0.91	−0.08	1.07	0.93
	100	−0.25	0.57	0.81	−0.26	0.72	0.84
	200	−0.30	0.44	0.75	−0.36	0.24	0.74

Note: See Note to Table 2. $w(y)$ denotes the weight function used in the estimating equation procedure.

5. Discussion

In this article, we have presented a general framework to the analysis of data from cancer studies using the proportional hazards model. More generally, this article provides a new way of thinking about such data. Provided one can make monotonicity assumptions such as those outlined in Kimmel and Flehinger (1991), one can analyze data on the time scale defined by the evolution of the tumor. Such assumptions can help provide information on the natural history of the disease as well as on the kinetics of progression. Novel features of the data not fully addressed before include the presence of length-biased sampling, right censoring, and the two-phase design of the study.

Another bias in screening studies is lead-time bias, in which the tumor is detected by screening but might potentially have no survival benefit. There are two potential extensions of this model to incorporate lead-time bias. One is to assume that lead-time bias detects all tumors at a size that is smaller than it would otherwise be. This would impose constraints on the form of w in (2). Another is to assume that lead-time bias would have no effect on tumor size of tumors that have metastasized but would detect a nonmetastatic tumor at a smaller tumor size than would be otherwise expected. For this situation, it would be possible to estimate the function w in (2). Both avenues are worthy of further exploration.

Throughout this article, we have assumed that the weight function in the length-bias sampling mechanism was known. It might be possible to estimate w if we change model assumptions, described at the beginning of this section, or if we attempted to incorporate tumor size/metastasis information from interval detected cases (cases detected by means other than screening between screening examinations). Without such data, the choice of w will typically depend on prior scientific or mechanistic considerations.

The estimation procedure described here for the proportional hazards regression model with right-censoring and length-biased sampling is novel. Proving asymptotic results about the estimators of the regression coefficients requires the use of modern empirical process techniques.

Finally, we have also shown that the Kimmel–Flehinger model for the Case I scenario is only valid under very restrictive assumptions, described in Section 2.3. One method of relaxing this assumption, suggested by a referee, is to allow ϕ to be subject specific, i.e., each subject has a specific growth function $\phi_i(t)$ $(i = 1, \ldots, n)$. To make estimation feasible, we will need to specify a model for $(\phi_1, \ldots, \phi_n)$, potentially as a function of covariates or by specifying a probability distribution for them. An alternative is to employ the Case II scenario from Ghosh (2006). However, this requires treating the tumor size data as interval-censored random variables that are subject to length-bias sampling, which is a much more complicated data structure.

6. Supplementary Materials

The Web Appendix referenced in Section 3.1 is available under the Paper Information link at the *Biometrics* website `http://www.tibs.org/biometrics`.

Acknowledgements

The author wishes to thank Jack Kalbfleisch and Jeremy Taylor for useful discussions. This research is supported in part by the National Institutes of Health through the University of Michigan's Cancer Center Support Grant (5 P30 CA46592).

References

Albert, A., Gertman, P. M., and Louis, T. A. (1978). Screening for the early detection of cancer—I. The temporal history of a progressive disease state. *Mathematical Biosciences* **40,** 1–59.

Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.

Asgharian, M. and Wolfson, D. B. (2005). Asymptotic behaviour of the unconditional NPMLE of the length-biased survivor function from right censored prevalent data. *Annals of Statistics* **33,** 2109–2131.

Baker, S. G., Erwin, D., and Kramer, B. S. (2003). Estimating the cumulative risk of a false positive cancer screenings. *BMC Medical Research Methodology* **3,** 11.

Barlow, W. E. (1994). Robust variance estimation of case–cohort design. *Biometrics* **50,** 1064–1072.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research, Volume 1*: *The Analysis of Case-Control Studies*. Lyon, France: IARC.

Cox, D. R. (1969). Some sampling problems in technology. In *New Development in Survey Sampling*, N. L. Johnson and H. Smith, Jr (eds), 506–527. New York: Wiley-Interscience.

Cox, D. R. (1972). Regression models and life tables (with Discussion). *Journal of the Royal Statistical Society, Series B* **34,** 187–220.

Day, N. E. and Walter, S. D. (1984). Simplified models of screening for chronic disease. Estimation procedure from mass screening programs. *Biometrics* **40,** 1–14.

Duchesne, T. and Lawless, J. (2000). Alternative time scales and failure time models. *Lifetime Data Analysis* **6,** 157–179.

Ghosh, D. (2006). Modeling tumor biology-progression relationships in screening trials. *Statistics in Medicine* **25,** 1872–1884.

Jones, M. C. (1991). Kernel density estimation for length biased data. *Biometrika* **78,** 511–519.

Kimmel, M. and Flehinger, B. J. (1991). Nonparametric estimation of the size–metastasis relationship in solid cancers. *Biometrics* **47,** 987–1004.

Lin, D. Y. (2000). On fitting Cox's proportional hazards models to survey data. *Biometrika* **87,** 37–47.

Lin, D. Y., Wei, L. J., Yang, I., and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, Series B* **62,** 711–730.

Lipsitz, S. R., Dear, K. B., and Zhao, L. P. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with application to clustered survival data. *Biometrics* **50,** 842–846.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, 2nd Edition.* New York: Wiley.

Patil, G. P. and Rao, C. R. (1978). Weighted distribution and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34,** 179–189.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology—Methodological Issues*, N. Jewell, K. Dietz, and V. Farewell (eds), 297–331. Boston: Birkhäuser.

Shen, Y. and Zelen, M. (1999). Parametric estimation procedures for screening programmes: Stable and nonstable disease models for multimodality case finding. *Biometrika* **86,** 503–515.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality.* New York: Springer Verlag.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *Annals of Statistics* **10,** 616–620.

Wang, M. C. (1996). Hazards regression analysis with length-biased data. *Biometrika* **83,** 343–354.

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

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

Zelen, M. and Feinleib, M. (1969). On the theory of screening for chronic diseases. *Biometrika* **56,** 601–614.

Received June 2006. *Revised February* 2007.
Accepted February 2007.