Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve

Author(s): Bradley Efron

Source: *Journal of the American Statistical Association*, Jun., 1988, Vol. 83, No. 402 (Jun., 1988), pp. 414-425

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association

Stable URL: https://www.jstor.org/stable/2288857

Logistic Regression, Survival Analysis, and the Kaplan–Meier Curve

BRADLEY EFRON*

We discuss the use of standard logistic regression techniques to estimate hazard rates and survival curves from censored data. These techniques allow the statistician to use parametric regression modeling on censored data in a flexible way that provides both estimates and standard errors. An example is given that demonstrates the increased structure that can be seen in a parametric analysis, as compared with the nonparametric Kaplan–Meier survival curves. In fact, the logistic regression estimates are closely related to Kaplan–Meier curves, and approach the Kaplan–Meier estimate as the number of parameters grows large.

KEY WORDS: Partial logistic regression; Parametric survival curves; Hazard-rate estimates; Parametric smoothing; Semiparametric smoothing.

1. INTRODUCTION

The Kaplan–Meier survival estimator is an important tool for analyzing censored data. Figure 1 shows a typical application. Two treatments for head and neck cancer were compared in a randomized trial. The Kaplan–Meier curves show treatment B outperforming treatment A in terms of patient survival, a result verified by standard significance tests. Good references for the Kaplan–Meier, or *product-limit* estimator, and life-table methods in general, include Miller (1981), Cox and Oakes (1984), Prentice and Kalbfleisch (1980), and Johnson and Elandt-Johnson (1980).

The Kaplan–Meier curve is so easy to calculate and (being totally nonparametric) requires so few assumptions that it is easy to forget its limitations. First of all, it can be inefficient compared to parametric survival estimators. This is the main point of Miller's important article "What Price Kaplan–Meier?" (Miller 1983). Secondly, survival curves are difficult to compare by eye, even in the absence of statistical noise.

Figure 2 compares treatments A and B for the cancer study in terms of their estimated hazard rates. These estimates are based on a parametric theory of survival-curve estimation, which is the main topic of this article. We now see quite a bit more structure. Both treatments start with the hazard rate near 0, followed by a high-risk period peaking at five months. The estimated hazard rates stabilize after one year, with treatment A having about 2.5 times higher risk than treatment B.

The method used to construct Figure 2, called *partial logistic regression,* is basically a straightforward application of logistic regression as described, for example, by Cox (1970). Section 2 gives an overview of this method, with the details filled in in Sections 3 and 4. All of this material involves a discretization of the data, even if it originally is in continuous form. Section 5 discusses the continuous limits of our discrete models. This clarifies their connection with traditional parametric survival functions such as the exponential and the Gompertz.

This article is primarily methodological. It shows how some familiar theoretical ideas (logistic regression, hazard-rate analysis, and partial likelihood) can be combined to give a simple, insightful analysis of censored data. Most of the theory has already appeared or is at least closely related to work by several authors: Cox (1975), Holford (1976), Thompson (1977), Efron (1977), Prentice and Gloeckler (1978), Pierce, Stewart, and Kopecky (1979), Anderson and Senthilselvan (1980, 1982), Padgett and Wei (1980), Mykytyn and Santner (1981), Laird and Oliver (1981), Tanner and Wong (1983, 1984), O'Sullivan (1986), Tsai (1986), and particularly Clayton (1983). These relationships are briefly discussed at the end of this article (Sec. 5, Remarks J and K).

2. PARTIAL LOGISTIC REGRESSION

This section discusses using traditional logistic regression techniques (see Cox 1970) to fit parametric survival curves to censored data. The method is easily implemented using any standard logistic regression program, gives direct estimates of the hazard rate, and provides approximate standard errors in addition to estimated survival curves and hazard rates. These parametric models are called *partial logistic regressions* because of a connection to Cox's (1975, ex. 2) theory of partial likelihood. The basic idea is not much different from that of the Kaplan–Meier estimator, and in fact reduces to the Kaplan–Meier estimator as the number of parameters grows large.

A concrete example will help explain the ideas and notation involved. Table 1 shows the data for arm A of the head-and-neck-cancer study, discretized by one-month intervals. The original undiscretized data appear at the bottom of the table. For each value of i from 1 to $N = 47$ (the last month with any data), the table shows

$$n_i = \text{no. of patients at risk at the beginning of month } i$$

$$s_i = \text{no. of patients who died during month } i$$

$$s_i' = \text{no. of patients lost to follow-up during month } i. \tag{2.1}$$

* Bradley Efron is Professor, Department of Statistics, Stanford University, Stanford, CA 94305. This article was stimulated by a talk of Wei-Yang Tsai concerning isotonic hazard-rate estimation.

Journal of the American Statistical Association
June 1988, Vol. 83, No. 402, Theory and Methods

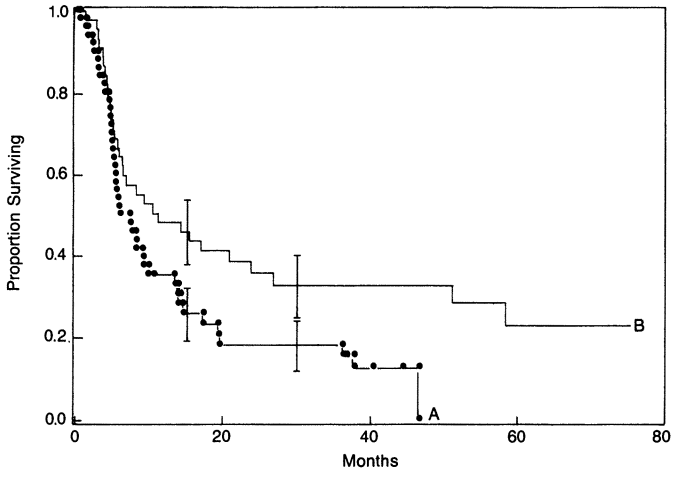

Figure 1. Kaplan–Meier Estimated Survival Curves, Arms A and B. These estimates are taken from a study comparing radiation therapy alone (A) versus radiation plus chemotherapy (B) for the treatment of head and neck cancer. Treatment B is significantly better according to the Mantel–Haenszel test, significance level .01 (see Tables 1 and 2). "Death" actually means "recurrence of disease." The error bars indicate ± one standard error.

For example, $n_3 = 48$ patients were alive at the beginning of the third month of observation, during which $s_3 = 5$ patients died and $s_3' = 1$ patient was lost to follow-up. This left $n_4 = 42$ patients still under study at the beginning of month 4. "Lost to follow-up," or "censored," data occurred mainly because patients entered the study at different calendar times, and some of them were still alive when the data were collected at the end of the study.

Table 2 shows the discretized data for arm B of the study. Here we have used $N = 61$ discrete intervals, not all of the same length. (The choice of discretization made little difference in the estimated hazard rates and survival curves; see Remark E, Sec. 3, and Remark I, Sec. 5.)

Our basic assumption is that for data of type (2.1), the number of deaths s_i is binomially distributed, given n_i, say

$$s_i \mid n_i \sim Bi(n_i, h_i) \quad \text{independently}, \qquad i = 1, 2, \ldots, N. \tag{2.2}$$

In other words, s_i has discrete density

$$\binom{n_i}{s_i} h_i^{s_i}(1 - h_i)^{n_i - s_i}, \qquad s_i = 0, 1, 2, \ldots, n_i.$$

Here h_i is the discrete hazard rate:

$$h_i = \Pr\{\text{patient dies during } i\text{th interval} \mid \text{patient survives until beginning of } i\text{th interval}\}. \tag{2.3}$$

The binomial assumption in (2.2) is basic to most work in survival analysis. Nice discussions appear in chapter 4 of Cox and Oakes (1984) and section (5.2) of Kalbfleisch and Prentice (1980). In what follows, *we consider the n_i to be fixed at their observed values, and take literally the independence assumption in* (2.2). Although this assumption cannot be exactly true (see Sec. 3, Remark A), it leads to reasonable conclusions under the usual assumptions for censored data.

The *survival function* for our discretized situation is

$$G_i \equiv \prod_{1 \le j < i} (1 - h_j), \tag{2.4}$$

the probability that a patient does *not* die during the first $i - 1$ time intervals and thus survives at least until the

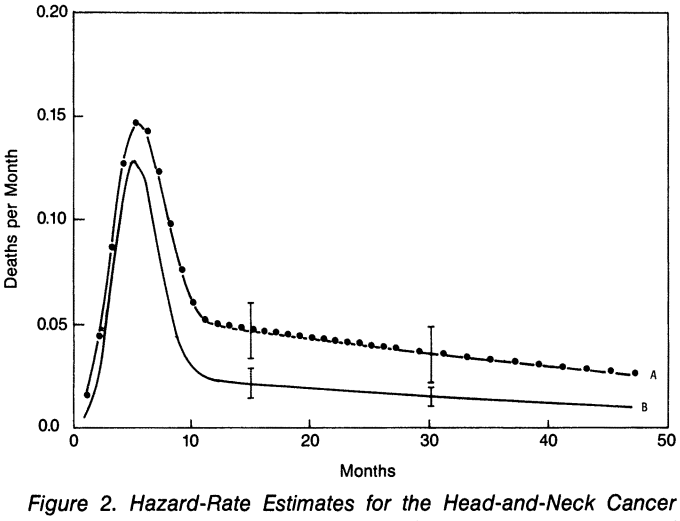

Figure 2. Hazard-Rate Estimates for the Head-and-Neck Cancer Study. There is an early high-risk period for both treatments. The hazard rates stabilize after one year, with treatment A having a hazard rate roughly 2.5 times that of treatment B. (The bullets are identifying symbols for curve A, not data points.) This figure is based on a parametric analysis described in Section 2.

Table 1. Data for Arm A of the Head-and-Neck-Cancer Study Conducted by the Northern California Oncology Group, Discretized by Months

Month	n_i	s_i	s_i'	Month	n_i	s_i	s_i'
1	51	1	0	25	7	0	0
2	50	2	0	26	7	0	0
3	48	5	1	27	7	0	0
4	42	2	0	28	7	0	0
5	40	8	0	29	7	0	0
6	32	7	0	30	7	0	0
7	25	0	1	31	7	0	0
8	24	3	0	32	7	0	0
9	21	2	0	33	7	0	0
10	19	2	1	34	7	0	0
11	16	0	1	35	7	0	0
12	15	0	0	36	7	0	0
13	15	0	0	37	7	1	1
14	15	3	0	38	5	1	0
15	12	1	0	39	4	0	0
16	11	0	0	40	4	0	0
17	11	0	0	41	4	0	1
18	11	1	1	42	3	0	0
19	9	0	0	43	3	0	0
20	9	2	0	44	3	0	0
21	7	0	0	45	3	0	1
22	7	0	0	46	2	0	0
23	7	0	0	47	2	1	1
24	7	0	0				
Total					628	42	9

NOTE: n_i is the number of patients at risk at the beginning of month i, s_i the number of observed deaths, s_i' the number lost to follow-up. The survival times in days for the 51 patients were 7, 34, 42, 63, 64, 74+, 83, 84, 91, 108, 112, 129, 133, 133, 139, 140, 140, 146, 149, 154, 157, 160, 160, 165, 173, 176, 185+, 218, 225, 241, 248, 273, 277, 279+, 297, 319+, 405, 417, 420, 440, 523, 523+, 583, 594, 1,101, 1,116+, 1,146, 1,226+, 1,349+, 1,412+, 1,417 ("+" indicates lost to follow-up). The table was constructed from these data, taking one month to be 30.438 days.

Table 2. Discretized Data for Arm B of the Head-and-Neck-Cancer Study

Month	n_i	s_i	s_i'	Month	n_i	s_i	s_i'
.5	45	0	0	23.0	14	0	0
1.0	45	0	0	24.0	14	1	0
1.5	45	1	0	25.0	13	0	1
2.0	44	0	0	26.0	12	0	0
2.5	44	0	0	27.0	12	1	0
3.0	44	1	0	29.0	11	0	0
3.5	43	2	0	31.0	11	0	0
4.0	41	3	0	33.0	11	0	0
4.5	38	3	0	35.0	11	0	0
5.0	35	2	0	37.0	11	0	1
5.5	33	2	0	39.0	10	0	0
6.0	31	2	1	41.0	10	0	1
6.5	28	2	0	43.0	9	0	0
7.0	26	1	0	45.0	9	0	1
7.5	25	0	0	47.0	8	0	0
8.0	25	0	0	49.0	8	0	0
8.5	25	1	0	51.0	8	0	0
9.0	24	0	0	53.0	8	1	0
10.0	24	1	0	55.0	7	0	1
11.0	23	1	0	57.0	6	0	0
12.0	22	1	0	59.0	6	1	1
13.0	21	0	0	61.0	4	0	0
14.0	21	0	0	63.0	4	0	1
15.0	21	1	0	65.0	3	0	0
16.0	20	1	0	67.0	3	0	1
17.0	19	0	0	69.0	2	0	0
18.0	19	1	2	71.0	2	0	1
19.0	16	0	0	73.0	1	0	0
20.0	16	0	0	75.0	1	0	0
21.0	16	1	1	77.0	1	0	1
22.0	14	0	0				
Total					1,123	31	14

NOTE: Discretization is by half-months until the end of month 9, by months until the end of month 27, and by two-month intervals until the end of month 77, for a total of $N = 61$ intervals. The survival time in days for the 45 patients were 37, 84, 92, 94, 110, 112, 119, 127, 130, 133, 140, 146, 155, 159, 169+, 173, 179, 194, 195, 209, 249, 281, 319, 339, 432, 469, 519, 528+, 547+, 613+, 633, 725, 759+, 817, 1,092+, 1,245+, 1,331+, 1,557, 1,642+, 1,771+, 1,776, 1,897+, 2,023+, 2,146+, 2,297+.

beginning of the ith interval ($G_1 = 1$, by definition). The life-table method estimates each h_i by $\tilde{h}_i = s_i/n_i$, the obvious choice from (2.2), and then substitutes $\tilde{h}_j$ for h_j in (2.4) to get the life-table survival estimate

$$\tilde{G}_i = \prod_{1 \le j < i} (1 - \tilde{h}_j). \tag{2.5}$$

The Kaplan–Meier curve is the same as (2.5), except that the time intervals are chosen so small that s_i never exceeds 1.

Our tactic in this paper is to estimate h_i by logistic regression. Let λ_i be the logistic parameter

$$\lambda_i \equiv \log[h_i/(1 - h_i)], \qquad i = 1, 2, \ldots, N, \tag{2.6}$$

so $h_i = [1 + \exp(-\lambda_i)]^{-1}$. For each value of i, let x_i be a known $1 \times p$ covariate vector. For example, we might consider a cubic logistic regression in time,

$$x_i = (1, t_i, t_i^2, t_i^3), \qquad i = 1, 2, \ldots, N, \tag{2.7}$$

where t_i is the midpoint of the ith time interval ($t_i = i - .5$ months in Table 1). The logistic regression model is

$$\lambda_i = x_i\alpha, \qquad i = 1, 2, \ldots, N, \tag{2.8}$$

where α is a $p \times 1$ vector of unknown parameters. A slightly more general form of (2.8) is used when the discretization intervals are of unequal lengths, as in Table 2. (See Sec. 3, Remark E.)

A standard logistic regression program—using the GLIM package, for example—finds the maximum likelihood estimate (MLE) $\hat{\alpha}$ for α, assuming $s_i \overset{\text{ind}}{\sim} Bi(n_i, h_i)$ as in (2.2). This gives $\hat{h}_i = [1 + \exp(-x_i\hat{\alpha})]^{-1}$ as the MLE of the hazard rate h_i, and $\hat{G}_i = \prod_{1<j<i} (1 - \hat{h}_j)$ as the MLE of the survival curve G_i. We call this procedure a partial logistic regression because of its connection with the theory of partial likelihood, see in particular example 2 of Cox (1975).

Figure 3 compares the life-table estimate $\tilde{G}_i$, (2.5), with two different partial logistic regression models. The triangles indicate $\hat{G}_i$, based on the cubic model (2.7). The bullets indicate $\hat{G}_i$ based on model (2.8), with $p = 4$ and

$$x_i = (1, t_i, (t_i - 11)_-^2, (t_i - 11)_-^3), \tag{2.9}$$

where $(t_i - 11)_- = \min\{(t_i - 11), 0\}$. Specification (2.9) is a cubic-linear spline, with the join point at $t = 11$ months. The logit λ_i is allowed to vary as a cubic function of time before $t = 11$ months, but only linearly in time after 11 months. Moreover, the logit thought of as a continuous function of time, say $\lambda(t) = \alpha_1 + \alpha_2 t + \alpha_3(t - 11)_-^2 + \alpha_4(t - 11)_-^3$, has an everywhere-continuous first derivative, even at $t = 11$. Figure 4 shows model (2.9) applied to arm B of the head-and-neck experiment.

The hazard-rate estimates in Figure 2 are based on the cubic-linear spline (2.9). The rationale for this model is simple: The head-and-neck-cancer study is typical of many medical survival situations in having more data available in the early months, and also in having a more complicated early structure. Model (2.9) is designed to match the complexity of the fitted curve to the availability of statistical information, and to the perceived need for complexity. Model (2.9), including the choice of the join point, is discussed further in Section 4.

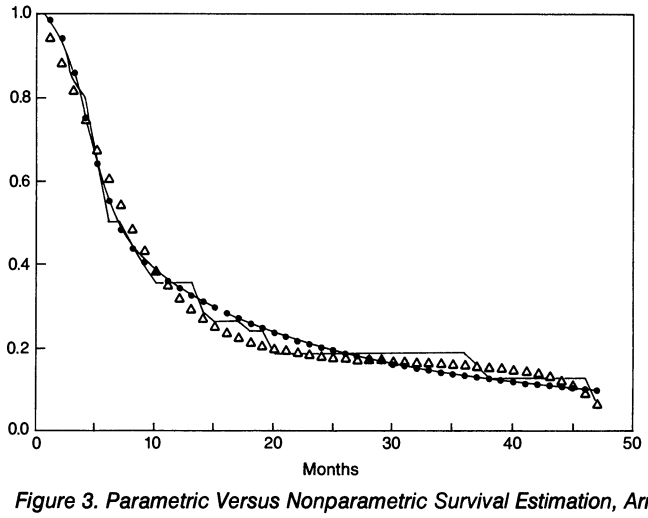

Figure 3. Parametric Versus Nonparametric Survival Estimation, Arm A. A life-table estimate for arm A (jagged solid line) is compared with two parametric estimates: cubic partial logistic regression (Δ) and cubic-linear spline joined at 11 months (smooth curve, indicated by ●).

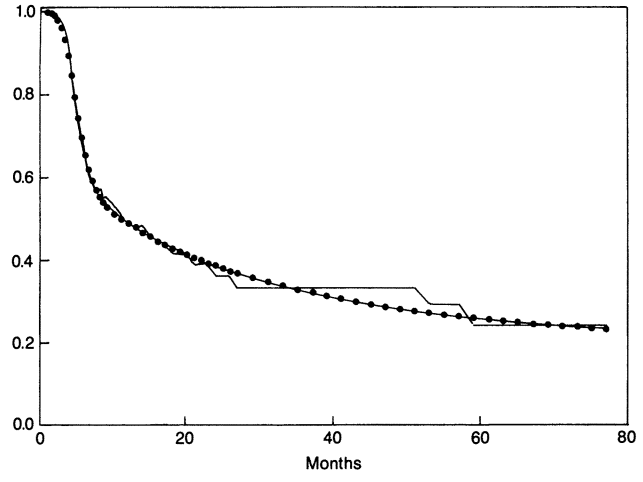

Figure 4. Parametric Versus Nonparametric Survival Estimates, Arm B. Life-table estimate for arm B (jagged line) is compared with partial logistic regression based on cubic-linear spline joined at 11 months [see (2.9)].

3. MAXIMUM LIKELIHOOD ESTIMATES AND STANDARD ERRORS

This section discusses calculation of maximum likelihood estimates and their standard errors, for partial logistic regression models. Using the arm A data of Table 1 as an example, we show how the parametric survival estimates approach the life-table curves and how their estimated standard errors approach those given by Greenwood's formula, as the parametric models become more complicated. The additional theory required for models involving join-point estimation, such as the cubic-linear spline (2.9), is discussed in Section 4.

Suppose, then, that we have $s_i \mid n_i \sim Bi(n_i, h_i)$ as in (2.2), where the n_i are considered fixed at their observed values, and the s_i are taken to be independently distributed, given the n_i. (The independence assumption is further discussed in Remark A.) Also, assume that the logistic parameter $\lambda_i = \log[h_i/(1 - h_i)]$ follows the linear logistic model $\lambda_i = x_i\alpha$ as in (2.8), so $h_i = [1 + \exp(-\lambda_i)]^{-1}$. We occasionally write $\lambda_{i,\alpha}$ and $h_{i,\alpha}$ to emphasize the dependence on α.

These assumptions describe a standard logistic regression model (e.g., see Cox 1970), so we will quote without proof the usual results for maximum likelihood estimation in such models. Let $\mathbf{s} = (s_1, s_2, \ldots, s_N)'$, $\mathbf{nh}_\alpha = (n_1h_{1,\alpha}, n_2h_{2,\alpha}, \ldots, n_Nh_{N,\alpha})'$, and X equal the $N \times p$ matrix having vector x_i of (2.8), as its ith row. Then the p-dimensional score vector $\dot{l}_\alpha = (\cdots (\partial/\partial\alpha_j) \log f_\alpha(\mathbf{s}) \cdots)'$ is

$$\dot{l}_\alpha = X'(\mathbf{s} - \mathbf{nh}_\alpha). \tag{3.1}$$

The MLE of α is that $\hat{\alpha}$ that makes (3.1) equal $\mathbf{0}$.

The $p \times p$ second derivative matrix $-\ddot{l}_\alpha$, with jlth element $-(\partial^2/\partial\alpha_j\partial\alpha_l) \log f_\alpha(\mathbf{s})$, is given by

$$-\ddot{l}_\alpha = X' \operatorname{diag}(n_iV_{i,\alpha})X. \tag{3.2}$$

Here $V_{i,\alpha} \equiv h_{i,\alpha}(1 - h_{i,\alpha})$, and $\operatorname{diag}(n_iV_{i,\alpha})$ is the $N \times N$ diagonal matrix with diagonal elements $n_iV_{i,\alpha}$. The *expected* Fisher information matrix for α, $\mathcal{I}_\alpha = E\{\dot{l}_\alpha(\mathbf{s})\dot{l}_\alpha(\mathbf{s})'\} = E\{-\ddot{l}_\alpha\}$, also equals $X' \operatorname{diag}(n_iV_{i,\alpha})X$. The *observed* Fisher information matrix is defined to be $\hat{\mathcal{I}} \equiv \mathcal{I}_{\hat{\alpha}}$, or equivalently from (3.2), $\hat{\mathcal{I}} = -\ddot{l}_{\hat{\alpha}} = X' \operatorname{diag}(n_iV_{i,\hat{\alpha}})X$.

Estimated standard errors (SE's) for quantities of interest such as $\hat{\alpha}$, $\hat{h}_i$, and $\hat{G}_i$ are obtained from familiar maximum likelihood calculations:

$$\begin{aligned}
\widehat{\operatorname{cov}}(\hat{\alpha}) &= \hat{\mathcal{I}}^{-1} \\
\widehat{\operatorname{SE}}(h_{i,\hat{\alpha}}) &= V_{i,\hat{\alpha}}[x_i\hat{\mathcal{I}}^{-1}x']^{1/2} \\
\widehat{\operatorname{SE}}(G_{i,\hat{\alpha}}) &= G_{i,\hat{\alpha}}\left[\left(\sum_{j<i} h_{j,\hat{\alpha}}x_j\right)\hat{\mathcal{I}}^{-1}\left(\sum_{j<i} h_{j,\hat{\alpha}}x_j\right)'\right]^{1/2}.
\end{aligned} \tag{3.3}$$

Here $G_{i,\hat{\alpha}} \equiv \prod_{j<i}(1 - h_{j,\hat{\alpha}})$. We usually use the shorter notation $\hat{h}_i = h_{i,\hat{\alpha}}$, $\hat{G}_i = G_{i,\hat{\alpha}}$.

Table 3 gives estimated hazard rates and their standard errors for three conditional logistic regressions (2.8), fit to the Table 1 arm A data: a linear model $x_i = (1, t_i)$, the cubic model (2.7), and the cubic-linear spline (2.9). (A

Table 3. Estimated Hazard Rates and Their Standard Errors at Selected Time Points, for Table 1 Arm A Data

	Hazard estimate					Standard error estimate			
Month	Linear	Cubic	Cubic-linear	Life-table	LTSM	Linear	Cubic	Cubic-linear	LTSM
1	.090	.053	.015	.020	.019	.0177	.0184	.0122	.0218
3	.085	.076	.087	.104	.071	.0152	.0165	.0228	.0160
5	.080	.095	.146	.200	.102	.0132	.0163	.0316	.0220
7	.076	.106	.123	.0	.099	.0117	.0191	.0319	.0197
9	.072	.107	.076	.095	.090	.0108	.0217	.0279	.0164
11	.068	.098	.051	.0	.075	.0102	.0218	.0185	.0199
15	.060	.068	.047	.083	.058	.0102	.0181	.0178	.0241
20	.052	.034	.043	.222	.038	.0113	.0132	.0131	.0225
25	.045	.016	.039	.0	.008	.0126	.0092	.0119	.0192
30	.039	.010	.035	.0	.0	.0136	.0072	.0133	.0162
35	.033	.010	.032	.0	.033	.0140	.0075	.0158	.0172
40	.029	.021	.029	.0	.053	.0145	.0138	.0183	.0193
45	.025	.112	.027	.0	.092	.0145	.0768	.0205	.0298
47	.023	.266	.026	.500	.099	.0140	.1956	.0213	.0331

NOTE: There are three conditional logistic regressions: linear, cubic, and cubic-linear spline, as explained in the text; the life-table estimate s_i/n_i; and a smoothed version of the life-table estimate (LTSM), obtained from an adaptive local regression smoother applied to the life-table estimate. Estimated standard errors for the linear and cubic models were obtained from (3.3). The standard error for cubic-linear spline includes a term for the choice of the join at 11 months, as explained in Section 4. The standard error for LTSM was obtained from 25 bootstrap replications.

Table 4. Residual Analysis of Four Hazard Models Fit to Table 1 Data

	Deaths		Expected Deaths				Signed deviance residuals			
Month	n	s	Linear	Cubic	Cubic-linear	LTSM	Linear	Cubic	Cubic-linear	LTSM
1	51	1	4.59	2.70	.76	.97	−2.10	−1.22	.27	.03
2	50	2	4.40	3.20	2.18	2.25	−1.33	−.74	−.13	−.17
3	48	5	4.08	3.65	4.16	3.41	.46	.70	.42	.84
4	42	2	3.49	3.65	5.31	3.82	−.90	−.98	−1.73	−1.07
5–6	72	15	5.70	7.06	10.40	7.41	3.44	2.78	1.46	2.63
7–8	49	3	3.68	5.24	5.41	4.78	−.38	−1.12	−1.19	−.91
9–11	56	4	3.93	5.77	3.54	4.63	.04	−.82	.25	−.31
12–14	45	3	2.88	3.81	2.18	2.89	.07	−.45	.54	.08
15–18	45	2	2.60	2.56	2.05	2.41	−.40	−.38	−.03	−.28
19–24	46	2	2.31	1.33	1.91	1.38	−.22	.55	.06	.50
25–31	49	0	2.02	.60	1.80	.84	−2.03	−1.09	−1.91	−.41
32–38	47	2	1.57	.49	1.52	1.45	.38	1.63	.38	.44
39–47	28	1	.75	1.96	.78	1.97	.28	−.78	.24	−.79
Total	628	42								
Sum of squares							23.71	18.24	11.02	11.04
Chi-squared significance level							.014	.032	.201	
Degrees of freedom							(11)	(9)	(8)	

NOTE: The models are the same ones considered in Table 3. The cubic-linear spline fits the data significantly better than either the linear or cubic models.

quadratic model was also used, but gave almost the same estimated hazards as the linear model.) The estimated standard errors for the linear and cubic regressions were obtained from (3.3). The standard error for the cubic-linear spline includes an additional term for estimating the join, as explained in Section 4. A semiparametric method (the smoothed life-table estimate) also appears in Table 3 and is discussed in Remark B of this section.

The linear model has the smallest estimated standard errors, but it does not fit the arm A data very well. This is seen in Table 4, where the data from Table 1 has been grouped into 13 time periods, each containing roughly 50 person-months of observation. For example, the seventh time period includes months 9–11, with 56 = 21 + 19 + 16 total person-months of observation and 4 = 2 + 2 + 0 total deaths, as shown in Table 1.

The expected deaths for each model,

$$e_j \equiv \sum_{\substack{j\text{th time} \\ \text{period}}} n_i \hat{h}_i, \tag{3.4}$$

appear in the middle panel of Table 4. How well or poorly these match the observed deaths s_j is measured by the *signed deviance residuals*

$$R_j \equiv \sqrt{2}\ \text{sign}(s_j - e_j) \times \left[s_j \log \frac{s_j}{e_j} + (n_j - s_j) \log \frac{n_j - s_j}{n_j - e_j} \right]^{1/2}, \tag{3.5}$$

shown in the right panel. If a model is correct—in the sense that it contains the true hazard function—then the R_j should be approximate standard normal deviates, with sum of squares Σh_j^2 approximately chi-squared distributed with 13 (no. of model parameters) df (see McCullagh and Nelder 1983, sec. 2.4.3). A numerical investigation suggested that replacing 13 by 14.3 increases the accuracy of the chi-squared approximation here, but does not substantially change the observed significance levels.

The bottom of the table shows significantly too-large values of ΣR_j^2 for the linear and cubic models, but the cubic-linear model, with 5 df including the choice of join, has an acceptable χ_8^2 significance level of .201. The differences in ΣR_j^2 are also significant; for example, 18.24 − 11.02 = 7.22 is .005 significant, compared with a χ_1^2 distribution, indicating a genuinely improved fit in going from a cubic to a cubic-linear model.

The nonparametric life-table estimate of the hazard rate is $\tilde{h}_i = s_i/n_i$. This estimate is always unbiased for h_i, assuming $n_i > 0$, but is usually too variable to be of direct use. (The excessive variability of $\tilde{h}_i$ is obvious in Table 3.) Of course, we can do better with a parametric model if the parametric assumptions are correct. The following results for conditional logistic regression models are further discussed in Remark D. The ratio of asymptotic variances of the parametric hazard estimate $\hat{h}_i = h_{i,\hat{\alpha}}$, compared with the nonparametric estimate $\tilde{h}_i$, is

$$\frac{\text{var}\{\hat{h}_i\}}{\text{var}\{\tilde{h}_i\}} = z_i \left[\sum_{j=1}^{N} z_j' z_j \right]^{-1} z_i', \qquad z_i \equiv \sqrt{n_i V_{i,\alpha}}\, x_i. \tag{3.6}$$

This ratio is always less than 1.

Theorem. The average ratio of asymptotic variances between the parametric and nonparametric hazard-rate estimates is

$$\frac{1}{N} \sum_{i=1}^{N} \frac{\text{var}\{\hat{h}_i\}}{\text{var}\{\tilde{h}_i\}} = \frac{p}{N}. \tag{3.7}$$

In other words, if we estimate the hazard rate using a p-parameter conditional logistic regression model, and if the model is correct, then the asymptotic variance of the hazard estimates is reduced by a factor p/N, compared with the nonparametric estimates, in the sense of (3.7). As $p \to N$ the advantage of $\hat{h}_i$ over $\tilde{h}_i$ disappears. In fact, if $p = N$ and the $N \times N$ matrix X is of full rank, then $\hat{h}_i = \tilde{h}_i$.

Table 5. Estimated Survival Functions and Standard Errors

	Estimated survival				Estimated standard errors for log survival			
Month	Linear	Cubic	Cubic-linear	Life-table	Linear	Cubic	Cubic-linear	Life-table
1	.910	.947	.985	.980	.019	.019	.031	.033
3	.759	.819	.860	.843	.054	.055	.059	.065
5	.640	.677	.642	.642	.084	.086	.090	.095
7	.545	.543	.483	.501	.109	.112	.129	.132
9	.469	.433	.402	.397	.131	.143	.155	.168
11	.407	.350	.359	.355	.150	.177	.186	.202
15	.313	.250	.295	.261	.184	.236	.256	.256
20	.236	.197	.235	.184	.222	.284	.286	.299
25	.185	.176	.191	.184	.262	.312	.309	.325
30	.150	.166	.159	.184	.307	.325	.319	.338
35	.125	.158	.134	.184	.358	.341	.324	.348
40	.107	.147	.115	.126	.413	.360	.338	.370
45	.094	.108	.100	.126	.470	.445	.436	.493
47	.089	.065	.095	.063	.493	.718	.686	.726

NOTE: Left panel: estimated survival $\hat{G}$ at the end of the indicated months, for four different estimators applied to the arm A data, Table 1. Right panel: estimated standard errors for $\log\{\hat{G}\}$. The standard error for the cubic-linear spline includes a term for the choice of join (see Sec. 4). Note that the life-table estimate is only slightly more variable than the cubic-linear spline.

Suppose we are interested in estimating the survival function G_i rather than the hazard rate h_i. In this case, parametric methods offer less impressive improvements over the nonparametric life-table approach. Table 5 compares the estimated survival curves $\hat{G}_i$ (from the linear, cubic, and cubic-linear spline models) with the nonparametric life-table estimate (2.5). The right panel shows estimated standard errors for $\log\{\hat{G}_i\}$. The cubic-linear spline, which was our only parametric model giving a satisfactory fit to the data in Table 1, is only slightly less variable than the life-table estimate. In the notation of (3.3),

$$\widehat{\mathrm{SE}}(\log G_{i,\hat{\alpha}}) = \left[\left(\sum_{j<i} h_{j,\hat{\alpha}} x_j\right) \hat{\mathcal{I}}^{-1} \left(\sum_{j<i} h_{j,\hat{\alpha}} x_j\right)'\right]^{1/2}. \quad (3.8)$$

It is easier to compare standard errors for $\log \hat{G}$ than for $\hat{G}$ itself, because the factor $G_{i,\hat{\alpha}}$ in the third equation of formula (3.3) is removed. To further sharpen the comparison, all of the standard errors in Table 5 were calculated assuming that the cubic model was true.

Formula (3.8) is closely related to Greenwood's formula for the variance of the life-table estimate. Suppose in (2.8) we take $p = N$ and $x_i = e_i$, the N-dimensional vector $(0, 0, \ldots, 1, 0, \ldots, 0)$ with 1 in the ith place ($i = 1, 2, \ldots, N$). Then X equals the $N \times N$ identity matrix, and (3.1) shows that the MLE $\hat{h}_i$ equals $s_i/n_i = \tilde{h}_i$, the nonparametric MLE. In this case, the MLE of the survival curve, $\hat{G}_i$, equals (2.5), the life-table estimate $\tilde{G}_i = \prod_{1<j<i}(1 - \tilde{h}_j)$. The observed information matrix $\hat{\mathcal{I}} = X'\,\mathrm{diag}(n_i V_{i,\hat{\alpha}})X$ equals $\mathrm{diag}(n_i\tilde{h}_i(1 - \tilde{h}_i))$, so (3.8) gives

$$\widehat{\mathrm{SE}}\{\log \tilde{G}_i\} = \left[\sum_{j<i} \frac{\tilde{h}_j^2}{n_j\tilde{h}_j(1 - \tilde{h}_j)}\right]^{1/2} = \left[\sum_{j<i} \frac{s_j}{n_j(n_j - s_j)}\right]^{1/2}, \quad (3.9)$$

which is Greenwood's formula (see Miller 1981, p. 45).

This calculation (as well as common sense) leads us to expect that the variability of a survival curve $\hat{G}_i$, obtained from a p-parameter conditional logistic regression, approaches the variability of the life-table estimate $\tilde{G}_i$ as $p \to N$. What is surprising in Table 5 is how quickly the approach takes place. Even the cubic model, with only $p = 4$ parameters, has barely 10%–15% smaller standard errors than $\tilde{G}_i$. On the other hand, parametric models provide much greater improvements when estimating the hazard rate, as the theorem (3.7) shows.

Remark A. The independence assumption (2.2) cannot be literally true. For example, if there is no censoring $s_i = n_i - n_{i+1}$. In this case, the sequence $s_1, s_2, \ldots,$ is completely determined by the sequence $n_1, n_2, \ldots,$ in contradiction to (2.2).

Nevertheless, calculations based on (2.2) give reasonable answers under reasonable assumptions. Using the notation of (2.1), let $\mathbf{v}_i' = (s_1, s_1', s_2, s_2', \ldots, s_{i-1}, s_{i-1}', s_i)$ and $\mathbf{v}_i = (s_1, s_1', s_2, s_2', \ldots, s_{i-1}, s_{i-1}')$. Starting with $n = n_1$ patients at risk at the beginning of observation (which we take to be a constant, fixed at its observed value), $\mathbf{v}_i$ is the history of deaths and losses for the first $i - 1$ time intervals; $\mathbf{v}_i'$ is the same history extended to include s_i. Here we follow the usual convention that the s_i' losses in any one time interval occur after the s_i deaths. Note that $n_2 = n_1 - s_1 - s_1'$, $n_3 = n_2 - s_2 - s_2'$, and so forth, so there is no need to indicate $n_2, n_3, \ldots, n_i$ in $\mathbf{v}_i$ or $\mathbf{v}_i'$.

We assume that s_i, given $\mathbf{v}_i$, has a $Bi(n_i, h_{i,\alpha})$ distribution, where $h_{i,\alpha} = [1 + \exp(-x_i\alpha)]^{-1}$, as in (2.8), and that s_i', given $\mathbf{v}_i'$, has a distribution depending on a nuisance parameter vector ξ, but not on α:

$$\begin{aligned} f_{\alpha,\xi}(s_1, s_1', s_2, s_2', \ldots, s_N, s_N') &= \left[\binom{n_1}{s_1} h_{1,\alpha}^{s_1}(1 - h_{1,\alpha})^{n_1 - s_1}\right] f_\xi(s_1' \mid \mathbf{v}_1') \\ &\quad \times \left[\binom{n_2}{s_2} h_{2,\alpha}^{s_2}(1 - h_{2,\alpha})^{n_2 - s_2}\right] f_\xi(s_2' \mid \mathbf{v}_2') \cdots \\ &\quad \times \left[\binom{n_N}{s_N} h_{N,\alpha}^{s_N}(1 - h_{N,\alpha})^{n_N - s_N}\right] f_\xi(s_N' \mid \mathbf{v}_N'). \end{aligned} \quad (3.10)$$

See Prentice and Kalbfleisch (1980, sec. 5.2) for a nice discussion of *noninformative censoring*, which (3.10) represents.

The log-likelihood $l_{\alpha,\xi} \equiv \log f_{\alpha,\xi}(s_1, s_1', \ldots, s_N')$ for (3.10) can be written as

$$l_{\alpha,\xi} = l_\alpha + l_\xi, \tag{3.11}$$

where

$$l_\alpha = \log \prod_{i=1}^{N} \binom{n_i}{s_i} h_{i,\alpha}^{s_i} (1 - h_{i,\alpha})^{n_i - s_i}$$

and l_ξ does not depend on α. Since l_α is the log-likelihood for the independent binomial situation $s_i \mid n_i \overset{\text{ind}}{\sim} Bi(n_i, h_{i,\alpha})$, we see that (a) the score vector for α based on (3.10) is $\dot{l}_\alpha = X'(\mathbf{s} - \mathbf{n}\mathbf{h}_\alpha)$, as given in (3.1); (b) the MLE $\hat{\alpha}$ is the same as that based on the independent binomial assumptions; (c) the second derivation matrix $-\ddot{l}_{\alpha,\xi}$ is of block-diagonal form,

$$-\ddot{l}_{\alpha,\xi} = \begin{pmatrix} -\ddot{l}_\alpha & 0 \\ 0 & -\ddot{l}_\xi \end{pmatrix}, \tag{3.12}$$

where $-\ddot{l}_\alpha$ is as given in (3.2); (d) the estimated covariance matrix for $\hat{\alpha}$, obtained by taking the upper left $p \times p$ block of $(-\ddot{l}_{\hat{\alpha},\xi})^{-1}$, is $-\ddot{l}_{\hat{\alpha}}^{-1} = \hat{\mathcal{I}}^{-1}$, where as before $\hat{\mathcal{I}}$ is the observed Fisher information matrix $X'\,\mathrm{diag}(n_i V_{i,\hat{\alpha}})X$, based on the independent binomial assumptions; (e) the expected Fisher information matrix $\mathcal{I}_{\alpha\xi} = E\{-\ddot{l}_{\alpha,\xi}\}$ is also of block-diagonal form, with the upper left $p \times p$ block equal to $E_{\alpha,\xi}\{-\ddot{l}_\alpha\}$; and (f) the approximate covariance matrix for $\hat{\alpha}$, obtained by taking the upper left $p \times p$ block of $\mathcal{I}_{\alpha\xi}^{-1}$, is $(E_{\alpha\xi}\{-\ddot{l}_\alpha\})^{-1}$, which by Jensen's inequality satisfies

$$(E_{\alpha\xi}\{-\ddot{l}_\alpha\})^{-1} \le E_{\alpha\xi}\{-\ddot{l}_\alpha^{-1}\}. \tag{3.13}$$

In summary, the independent binomial model (2.2)–(2.8) can be replaced by the more believable model (3.10), without changing the MLE $\hat{\alpha}$. The estimated covariance matrix of $\hat{\alpha}$, based on the observed Fisher information, also does not change. This is not true for the expected Fisher information, but (3.13) suggests that the covariance approximation based on the independent binomial model will be greater than that based on (3.10).

Remark B. The smoothed life-table hazard estimate (LTSM) in Table 3 was obtained by applying an adaptive local regression smoother ("super-smoother"; see Friedman and Stuetzle 1981) to the life-table estimates $\tilde{h}_i = s_i/n_i$. "Adaptive" here means that the width of the local smoothing window was chosen by the smoother itself, on the basis of a generalized cross-validation criterion. Hastie and Tibshirani (1986) discussed local smoothing estimators, extensively.

Table 3 shows only moderate agreement between LTSM and the cubic-linear spline hazard estimate for arm A. The single death of 47 months in Table 1 greatly influenced LTSM. Removing this death brings LTSM down to 0 at 47 months. The agreement between LTSM and the cubic-linear spline was excellent for arm B.

Using a model such as a cubic-linear spline can be seen as a strict form of smoothing. Past 15 months, the spline model has a considerably smaller standard error than LTSM (see Table 3), but may be biased if the model is grossly wrong. Remark H of Section 5 briefly discusses the literature of smooth hazard estimators.

Remark C. The referees for this article suggested using the *complementary log-log link* $\phi_{i,\alpha} = \log(-\log(1 - h_{i,\alpha}))$ instead of the logit $\lambda_{i,\alpha}$ as our regression parameter. For small values of $h_{i,\alpha}$ such as those suggested by Tables 1 and 2, this makes little difference, since $\lambda_{i,\alpha}$ and $\phi_{i,\alpha}$ are nearly equal. As a check, the cubic-linear spline model was refit, using $\phi_{i,\alpha} = x_i\alpha$ instead of (2.8). The refitted hazard rate $h_{i,\hat{\alpha}}$ for arms A and B agreed with the cubic-linear column in Table 3 to within .3% for every value of t.

Remark D. Let $\hat{\mathbf{h}} = (h_{i,\hat{\alpha}}, h_{2,\hat{\alpha}}, \ldots, h_{N,\hat{\alpha}})'$ be the vector of estimated hazards based on a partial logistic regression (2.2) and (2.8), and let $\tilde{\mathbf{h}} = (\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_N)'$ be the vector of life-table hazards $\tilde{h}_i = s_i/n_i$. Also, let $\mathcal{I}_\alpha = X'\,\mathrm{diag}(n_i V_{i,\alpha})X$, as defined after (3.2). The joint covariance matrix of $\hat{\mathbf{h}}$ and $\tilde{\mathbf{h}}$ is approximately

$$\mathrm{cov}\begin{pmatrix} \hat{\mathbf{h}} \\ \tilde{\mathbf{h}} \end{pmatrix} \doteq \begin{pmatrix} M_1 & M_1 \\ M_1 & M_2 \end{pmatrix}, \tag{3.14}$$

where M_1 and M_2 are $N \times N$ matrices: $M_1 = \mathrm{diag}(V_{i,\alpha})X\mathcal{I}_\alpha^{-1}X'\,\mathrm{diag}(V_{i,\alpha})$ and $M_2 = \mathrm{diag}(V_{i,\alpha}/n_i)$. [This follows from the Taylor series expansion $\hat{\mathbf{h}} - \mathbf{h} \doteq \mathrm{diag}(V_{i,\alpha})X\mathcal{I}_\alpha^{-1}X'(\mathbf{s} - n\mathbf{h}_\alpha)$ and $\tilde{\mathbf{h}} - \mathbf{h} = \mathrm{diag}(1/n_i)(\mathbf{s} - n\mathbf{h}_\alpha)$.]

The form of (3.14) shows that, asymptotically, $\tilde{\mathbf{h}} = \hat{\mathbf{h}}$ + independent extra error; the covariance of the extra error is $M_2 - M_1$, which is a positive semidefinite matrix. Note that (3.14) gives (3.6). Then,

$$\begin{aligned} &\frac{1}{N}\sum_{i=1}^{N} \frac{\mathrm{var}\{\hat{h}_i\}}{\mathrm{var}\{\tilde{h}_i\}} \\ &\quad \doteq \frac{1}{N}\sum_{i=1}^{N} z_{i,\alpha}\left[\sum_j z_{j,\alpha}' z_{j,\alpha}\right]^{-1} z_{i,\alpha}' = \frac{1}{N}\,\mathrm{tr}\{I_p\} \\ &\quad = p/N, \end{aligned} \tag{3.15}$$

which is the theorem (3.7). The asymptotic independence of $\tilde{\mathbf{h}} - \hat{\mathbf{h}}$ and $\hat{\mathbf{h}}$ is a familiar result from general theory relating efficient estimators and unbiased estimates of 0 (see Rao 1973, sec. 5a); some version of (3.15) would hold for parameterizations other than the logistic regression (2.8).

Remark E. A more general form of the linear logistic model (2.8) is

$$\lambda_i = a_i + x_i\alpha, \qquad i = 1, 2, \ldots, N, \tag{3.16}$$

where $a_1, a_2, \ldots, a_N$ are any fixed constants. Results (3.1)–(3.3) remain valid as stated, as long as we remember that λ_i is given by (3.16) rather than (2.8). McCullagh and Nelder (1984, p. 138) called a_i an *offset*.

The offset form of $\lambda_{i,\alpha}$ was used in fitting conditional

logistic regression models to the arm B data of Table 2:

$$\begin{aligned} a_i &= \log(\tfrac{1}{2}), \quad i = 1, \ldots, 18 \\ &= \log(1), \quad i = 19, \ldots, 36 \\ &= \log(2), \quad i = 37, \ldots, 61. \end{aligned} \tag{3.17}$$

These a_i compensate for the differing lengths of the time intervals in Table 2. [See Eq. (5.3). The fitted hazards $h_{i,\hat{\alpha}}$ for arm B were adjusted to one-month intervals, for example, by multiplying $h_{i,\hat{\alpha}}$ by 2 for $i = 1, \ldots, 18$, in order to make the plotted hazard rates in Fig. 2 comparable.]

4. THE CUBIC-LINEAR SPLINE MODEL

This section discusses the maximum likelihood estimation of the join point in the cubic-linear spline model (2.9). We are particularly interested in assessing the increased standard error of quantities of interest, such as the hazard-rate differences between arms A and B of the cancer study, due to the estimation of the join. The discussion here is very brief. Efron (1986, sec. 4) gave more details.

Tables 6 and 7 numerically summarize the rather technical results of Efron (1986). In Table 6 we see the MLE's of the hazard rates for arms A and B, based on the cubic-linear spline model with join at 11 months. The estimated standard errors are calculated from a formula (4.3) that adds a term to (3.3) to account for the data-based selection of the join. The *penalty ratio* {standard error from (4.3)}/{standard error from (3.3)} can be quite substantial, the greatest penalty in Table 6 being 1.48.

Table 7 compares the hazard rates for arms A and B. The two comparison statistics are the difference of the hazard rates, say $\hat{h}_{A,i} - \hat{h}_{B,i}$, and the difference on the logit scale $\hat{\lambda}_{A,i} - \hat{\lambda}_{B,i} = \log(\hat{h}_{A,i}/\hat{h}_{B,i}) - \log((1 - \hat{h}_{A,i})/(1 - \hat{h}_{B,i}))$. In this case, we see that there is almost no penalty from estimating the join.

Tables 6 and 7 have a simple interpretation: Choosing the join point on the basis of the data can substantially increase the standard error of the estimated hazard rates but it has little effect on statistics comparing the two hazard rates. In Figure 2, for instance, we can have greater faith in the pictured differences of the two hazards than their individual standard errors would suggest. (It is important to note that this statement depends on choosing the *same* join for both estimated hazard rates.)

Table 6. Estimated Hazard Rates, Their Standard Errors, and the Penalty Ratio at Selected Time Points

	Arm A estimates			Arm B estimates		
Month	$\hat{h}$	se	Penalty	$\hat{h}$	se	Penalty
1	.015	.012	1.11	.004	.005	1.17
3	.087	.023	1.06	.060	.021	1.02
5	.146	.032	1.02	.129	.035	1.03
7	.123	.032	1.29	.091	.030	1.33
9	.076	.028	1.38	.040	.019	1.48
11	.051	.018	1.04	.024	.009	1.00
15	.047	.018	1.30	.021	.009	1.22
20	.043	.013	1.17	.019	.007	1.18
25	.039	.012	1.03	.017	.006	1.11
35	.032	.016	1.02	.013	.005	1.10
45	.027	.021	1.06	.010	.005	1.01

NOTE: The penalty ratio equals {se including join choice}/{se join prechosen}. The penalty ratio can be quite large.

Table 7. Estimated Differences Between Arms A and B

	Logit difference			Hazard difference		
Month	$\hat{\lambda}_A - \hat{\lambda}_B$	se	Penalty	$\hat{h}_A - \hat{h}_B$	se	Penalty
1	1.31	1.41	1.03	.011	.012	1.00
3	.43	.45	1.00	.022	.030	1.00
5	.21	.37	1.00	.017	.044	1.00
7	.39	.36	1.02	.032	.033	1.00
9	.68	.45	1.04	.035	.024	1.00
11	.80	.54	1.00	.028	.020	1.01
15	.83	.48	1.00	.026	.016	1.02
20	.85	.42	1.01	.024	.013	1.01
25	.87	.44	1.01	.022	.013	1.00
35	.90	.63	1.01	.019	.016	1.01
45	.95	.92	1.01	.016	.020	1.02

NOTE: Left panel: the logit scale. Right panel: the hazard scale. The penalty ratio is now very small, so estimating the join point from the data adds little to the standard error.

The results in Table 6 are based on a generalization of model (2.8):

$$\lambda_i \equiv \lambda_{i,\alpha,\phi} = x_i(\phi)\alpha, \qquad i = 1, 2, \ldots, N, \tag{4.1}$$

where $x_i(\phi)$ is a $1 \times p$ covariate vector depending on an unknown real-valued parameter ϕ. We require that the vector derivative

$$\dot{x}_i(\phi) \equiv \left(\cdots \frac{dx_{ij}(\phi)}{d\phi} \cdots \right) \tag{4.2}$$

be defined continuously as a function of ϕ.

The case of particular interest is where ϕ is the join point of a cubic-linear spline, $x_i(\phi) = (1, t_i, (t_i - \phi)_-^2, (t_i - \phi)_-^3)$, as in (2.9). Then $\dot{x}_i(\phi) = (0, 0, -2(t_i - \phi)_-, -3(t_i - \phi)_-^2)$, a continuous function of ϕ for any value t_i.

Suppose that α and ϕ in (4.1) are estimated by maximum likelihood. Then the estimated hazard rate $h_{i,\hat{\alpha},\hat{\phi}} = [1 + \exp(-\lambda_{i,\hat{\alpha},\hat{\phi}})]^{-1}$ has greater standard error than indicated in (3.3) because of the estimation of ϕ in addition to α, and likewise for the other estimated parameters. The appropriate standard-error formulas require the following definitions: X and $\dot{X}$ are the $N \times p$ matrices, with ith rows $x_i(\phi)$ and $\dot{x}_i(\phi)$, respectively; $D = \operatorname{diag}(n_i V_{i,\hat{\alpha},\hat{\phi}})$, where $V_{i,\alpha,\phi} = h_{i,\alpha,\phi}(1 - h_{i,\alpha,\phi})$; $\hat{\mathcal{I}} = X'DX$; $v = \hat{\mathcal{I}}^{-1}(X'D\dot{X}\hat{\alpha})$; $\mathbf{u} = \dot{X}\hat{\alpha} - Xv$; and $Q = \sum_{j=1}^{N} u_j^2 n_j V_{j,\hat{\alpha},\hat{\phi}}$.

The asymptotic formulas for the standard errors are

$$\begin{aligned} \widehat{\mathrm{SE}}(\hat{\lambda}_i) &\doteq [x_i \hat{\mathcal{I}}^{-1} x_i' + u_i^2/Q]^{1/2} \\ \widehat{\mathrm{SE}}(\hat{h}_i) &\doteq V_{i,\hat{\alpha},\hat{\phi}}[x_i \hat{\mathcal{I}}^{-1} x_i' + u_i^2/Q]^{1/2} \\ \widehat{\mathrm{SE}}(\hat{G}_i) &\doteq G_{i,\hat{\alpha},\hat{\phi}} \left[\left(\sum_{j<i} h_j x_j \right) \hat{\mathcal{I}}^{-1} \left(\sum_{j<i} h_j x_j \right)' \right. \\ &\qquad \left. + \left(\sum_{j<i} h_j u_j \right)^2 \Big/ Q \right]^{1/2}. \end{aligned} \tag{4.3}$$

Table 8. Deviances as a Function of the Join Point $\hat{\phi}$ for Arms A and B

	$\hat{\phi}$				
Deviance	*9.5*	*10*	*11*	*12*	*13*
dev_A	50.928	48.016	47.519	47.370*	47.669
dev_B	34.011*	34.083	34.557	35.205	35.889
Total	84.939	82.099	82.076*	82.575	83.558

NOTE: The total deviance is minimized for $\hat{\phi} = 11$.
* Minima.

Comparing (4.3) with (3.3), we see that the terms in (4.3) involving u_i represent the additional standard error from estimating ϕ by its MLE $\hat{\phi}$. These terms disappear if ϕ is assumed known, in which case the estimated standard errors from (4.3) are the same as those from (3.3).

Efron (1986, sec. 4), derived formulas (4.3), in addition to more complicated formulas for the standard errors of quantities such as $\hat{h}_{A,i} - \hat{h}_{B,i}$, which compare two estimated hazard rates. Table 7 shows that these complicated formulas are almost superfluous here. We get almost the same estimated standard errors, simply by using (3.3) on the two independent arms of the study: $\widehat{\text{SE}}(\hat{h}_{A,i} - \hat{h}_{B,i}) = [\widehat{\text{SE}}(\hat{h}_{A,i})^2 + \widehat{\text{SE}}(\hat{h}_{B,i})^2]^{1/2}$. In other words, in this case we can ignore the effect of estimating ϕ.

Remark F. The MLE's $(\hat{\alpha}_A, \hat{\alpha}_B, \hat{\phi})$ for the cancer study were found by combining standard logistic regression for $\hat{\alpha}_A$ and $\hat{\alpha}_B$ with a direct search over the possible values of ϕ. Table 8 shows part of this search. For each trial value of $\hat{\phi}$, the MLE's $\hat{\alpha}_A$ and $\hat{\alpha}_B$ were found with a logistic regression program, producing estimates $\hat{h}_{A,i}$ and $\hat{h}_{B,j}$. The deviances dev_A and dev_B,

$$\text{dev} \equiv 2\sum_{i=1}^{N}\left(s_i \log\frac{s_i}{n_i\hat{h}_i} + (n_i - s_i)\log\frac{n_i - s_i}{n_i(1-\hat{h}_i)}\right),$$

get smaller as the likelihood gets larger. The value of $\hat{\phi}$ that minimizes $\text{dev}_A + \text{dev}_B$ is the MLE. Table 8 shows that $\hat{\phi} = 11$ for the cancer study.

5. THE CONTINUOUS CASE

Our methods so far have depended on discretization of the data, as in Tables 1 and 2. This introduces an arbitrary feature into the analysis, although the exact form of the discretization is usually unimportant (see Remark I). This section discusses the class of continuous models obtained as limits of the conditional logistic regressions (2.2) and (2.8), as the discrete time intervals ("months" in Table 1) become small. The discussion by Clayton (1983) is closely related to the point of view taken here.

Suppose that $g_\alpha(t)$ is a probability density function on the positive axis, with survival function $G_\alpha(t) = \int_t^\infty g_\alpha(s)\,ds$ and hazard function $h_\alpha(t) = g_\alpha(t)/G(t)$ for $t > 0$. It is convenient to assume that $g_\alpha(t)$ has a continuous second derivative in t, though only a continuous first derivative is actually necessary to make the connection with the earlier sections.

It is easy to see the results of discretizing this continuous situation. Suppose that the ith discrete time interval has center point t_i and length Δ_i. Then, the discrete density $g_{i,\alpha} = \int_{t_i-\Delta_i/2}^{t_i+\Delta_i/2} g_\alpha(t)\,dt$ is obtained by a standard Taylor series argument:

$$g_{i,\alpha} = g_\alpha(t_i)\Delta_i + O(\Delta_i^3). \tag{5.1}$$

Similarly, the discrete survival function $G_{i,\alpha} = \sum_{j\geq i} g_{i,\alpha}$ and discrete hazard rate $h_{i,\alpha} = g_{i,\alpha}/G_{i,\alpha}$ are given by

$$G_{i,\alpha} = G_\alpha(t_i) + [g_\alpha(t_i)/2]\,\Delta_i + O(\Delta_i^2)$$

$$h_{i,\alpha} = h_\alpha(t_i)\,\Delta_i - \tfrac{1}{2}h_\alpha(t_i)^2\,\Delta_i^2 + O(\Delta_i^3). \tag{5.2}$$

The logistic parameter $\lambda_{i,\alpha} = \log h_{i,\alpha}/(1 - h_{i,\alpha})$ equals

$$\lambda_{i,\alpha} = \log(\Delta_i) + \log(h_\alpha(t_i)) + \tfrac{1}{2}h_\alpha(t_i)\Delta_i + O(\Delta_i^2). \tag{5.3}$$

Note that the parameter $\phi_{i,\alpha} = \log(-\log(1 - h_{i,\alpha}))$ discretizes more nicely. Thus $\phi_{i,\alpha} = \log(\Delta_i) + \log(h_\alpha(t_i)) + O(\Delta_i^3)$, but as Remark C shows, this is unimportant in the present context.

In this section we consider the parametric class of continuous hazard functions

$$h_\alpha(t) = \exp[x(t)\alpha], \tag{5.4}$$

where α is a $p \times 1$ vector of unknown parameters and $x(t)$ is an observed p-dimensional time-dependent covariate vector. For convenience, we assume that $x(t)$ has first coordinate 1 for all t, say $x(t) = (1, x_{(1)}(t))$, and that $x_{(1)}(t)$ has continuous second derivatives with respect to t. If all of the discrete time intervals have the same length $\Delta_i = \Delta$, then (5.3) gives

$$\lambda_{i,\alpha} = \log(\Delta) + x(t_i)\alpha + O(\Delta). \tag{5.5}$$

This approaches model (2.8) as $\Delta \to 0$, with the quantity $\log(\Delta)$ in (5.5) absorbed into the α_1 term of $x(t_i)\alpha = \alpha_1 + x_{(1)}(t_i)\alpha_{(1)}$.

The simplest case of (5.4) is where $p = 1$ and $x(t) = 1$, so $h_\alpha(t) = \exp(\alpha_1)$ for all $t > 0$. Let $\theta \equiv \exp(\alpha_1)$. The *cumulative hazard* $H_\alpha(t) \equiv \int_0^t h_\alpha(s)\,ds$ equals θt, giving $G_\alpha(t) = \exp[-H_\alpha(t)] = \exp(-\theta t)$ and $g_\alpha(t) = \exp(-\theta t)$ for $t > 0$. In other words, the lifetime distribution T corresponding to hazard function $\exp(\alpha_1) = \theta$ is a one-sided exponential distribution scaled to have expectation θ^{-1}, say $T \sim Z/\theta$, where Z has density e^{-z} for $z > 0$.

The next-simplest case is where the log hazard rate is linear in t, $x(t) = (1, t)$, so that

$$h_\alpha(t) = \exp(\alpha_1 + \alpha_2 t), \qquad t > 0. \tag{5.6}$$

Defining $\theta \equiv [\exp(\alpha_1)]/\alpha_2$, the cumulative hazard and survival functions are

$$H_\alpha(t) = \theta[\exp(\alpha_2 t) - 1]$$

$$G_\alpha(t) = \exp - \{\theta[\exp(\alpha_2 t) - 1]\}. \tag{5.7}$$

We recognize $G_\alpha(t)$ as the survival function of a Gompertz distribution (see Johnson and Kotz 1970, p. 271). In other words, (5.6) corresponds to a Gompertz lifetime distribution,

$$T \sim (1/\alpha_2)\log[1 + Z/\theta], \tag{5.8}$$

where Z is a standard one-sided exponential. This last result assumes that $\alpha_2 > 0$, so $G_\alpha(t)$ approaches 0 as $t \to \infty$.

What if $\alpha_2 < 0$ in (5.6)? This represents the interesting case of an *incomplete* Gompertz distribution, where there is positive probability that the lifetime T is infinite:

$$P_\alpha\{T = \infty\} = e^\theta, \qquad \theta \equiv [\exp(\alpha_1)]/\alpha_2 < 0. \quad (5.9)$$

With this understanding, (5.7) remains true as stated.

All of the discrete models considered previously have the potential for estimating $P_\alpha\{T = \infty\}$ to be positive. (This happened in both arms of the cancer study; see Remark G.) This is an advantage of our approach. In practice, it is often desirable to include the possibility of a positive *survival fraction,* but this can be clumsy to do with the usual parametric models for lifetime distributions. Miller (1981, sec. 2.4) gave a brief discussion.

More complicated examples of model (5.4), such as log $h_\alpha(t) = \alpha_1 + \alpha_2 t + \alpha_3 t^2$, do not yield simple expressions for the cdf or density. This is unimportant, since the estimation of parameters depends only on the log hazard rate, which is particularly easy to use for model (5.4).

The parameter vector α in model (5.4) is estimated as follows: Let $n(t)$ be the number of patients at risk just before time t. We assume that the occurrence of observed deaths is a Poisson point process, with intensity $n(t)h_\alpha(t) = n(t)e^{x(t)\alpha}$ at time t. This is the limiting process obtained from (2.2) by letting the discrete time intervals decrease to zero length (see Efron 1977). Suppose that out of all n patients we observed m deaths, at times, say, $T_1, T_2, \ldots, T_m$, with the other $n - m$ patients being lost to follow-up at various times during the study. Define $S \equiv \sum_{j=1}^m x(T_j)'$. The score vector $\dot{l}_\alpha$ for the Poisson process is

$$\dot{l}_\alpha = S - \int_0^\infty n(t)x(t)'e^{x(t)\alpha}\,dt, \quad (5.10)$$

so the MLE $\hat{\alpha}$ is given by

$$S = \int_0^\infty n(t)x(t)'e^{x(t)\hat{\alpha}}\,dt. \quad (5.11)$$

The observed Fisher information matrix for α is

$$-\ddot{l}_\alpha = \int_0^\infty n(t)x(t)'x(t)e^{x(t)\alpha}\,dt. \quad (5.12)$$

It is easy to see that (5.10) and (5.12) are simply the continuous analogs of (3.1) and (3.2). Conversely, one can look at (3.1) and (3.2) as convenient summation approximations to the integrals in (5.10)–(5.12). The connection between the discrete and continuous cases was drawn more carefully in Efron (1977), including a derivation of (5.10)–(5.12).

Remark G. A continuous cubic-linear spline model log $h_\alpha(t) = \alpha_1 + \alpha_2 t + \alpha_3(t - \phi)_-^2 + \alpha_4(t - \phi)_-^3$ has

$$\log h_\alpha(t_1) = \log h_\alpha(t_0) + \alpha_2(t_1 - t_0) \quad (5.13)$$

for values of t_1 and t_0 greater than the join point ϕ. Calculations such as those for the Gompertz distribution (5.7) give

$$G_\alpha(t_1)/G_\alpha(t_0) = \frac{h_\alpha(t_0)}{-\alpha_2}\{1 - \exp[\alpha_2(t_1 - t_0)]\},$$
$$t_1 \geq t_0 \geq \phi. \quad (5.14)$$

If $\alpha_2 < 0$, then $P_\alpha\{T = \infty\}$ is positive and can be found by letting $t_1 \to \infty$ in (5.14):

$$P_\alpha\{T = \infty\} = G_\alpha(t_0)[h_\alpha(t_0)/(-\alpha_2)]. \quad (5.15)$$

For both arms of the cancer study, the MLE $\hat{\alpha}_2$ was negative. Formula (5.15), with $t_0 = 47$ months for arm A and $t_0 = 77$ months for arm B, gives the following estimates for the survival fractions:

$$P_{\hat{\alpha}_A}\{T = \infty\} = .025, \qquad P_{\hat{\alpha}_B}\{T = \infty\} = .189. \quad (5.16)$$

Of course, estimates such as (5.16) should be interpreted with caution, since they represent heroic extrapolations beyond the observed data.

Remark H. This article concentrates on the one-sample situation, where all patients have the same survival curve $G_\alpha(t)$. Model (5.4) and its discrete analog extend easily to the regression situation, where patient j's survival depends on a time-varying vector $z_j(t)$ of observed covariates, say

$$h_j(t) = \exp[x(t)\alpha + z_j(t)\beta]. \quad (5.17)$$

Model (5.17), and in particular its connection with Cox's *partial likelihood,* or *proportional hazards,* model was examined in Efron (1977). It is shown there that the fully parametric model (5.17) will usually not improve much on the partial likelihood model $h_j(t) = h_0(t)\exp[z_j(t)\beta]$, with $h_0(t)$ completely unspecified, at least not for the estimation of β. On the other hand, (5.17) can be effective in actually estimating the hazards $h_j(t)$, rather than just comparing them as the partial likelihood model does.

Remark I. Suppose that in the continuous Poisson-process situation (5.4), we discretize to situation (2.2). How much information is lost? For convenience assume that the continuous lifetime variate T takes its values in the unit interval [0, 1], and that the discretization of the data is into N equal subintervals, as in Table 1. Let $\mathcal{I}_\alpha(N)$ be the Fisher information matrix for α based on the discrete data (2.2) (taking the independence assumption literally), and let $\mathcal{I}_\alpha(\infty)$ be the Fisher information matrix based on the original continuous data. Then, as $N \to \infty$,

$$\mathcal{I}_\alpha(N) \doteq \mathcal{I}_\alpha(\infty) - c/N^2, \quad (5.18)$$

where $c = (1/12)\int_0^1 \dot{x}(t)'\dot{x}(t)n(t)h(t)\,dt$. Here the function $n(t)$ is considered fixed at its observed value, even though it is random [like the n_i in (2.2)].

Result (5.18) says that the information loss due to discretization goes to 0 very quickly as N grows large. Various alternative discretizations were tried on the cancer-study data, such as discretizing arm A into the same intervals used for arm B in Table 2, with almost-imperceptible changes in the results.

Remark J. A data discretization different from (2.2) was used by Holford (1976) and Laird and Oliver (1981): Assume that the lifetime variate T is "piecewise exponential," that is, that T has a constant hazard rate within each discrete time interval. Let θ_i be the hazard rate for the ith interval, and let r_i be the total observation time for all patients during the ith interval.

The likelihood function for this situation is the same as that for a model which assumes that the number of deaths s_i in interval i is conditionally Poisson,

$$s_i|r_i \sim Po(\theta_i r_i) \quad \text{independently}, \qquad i = 1, 2, \ldots, N. \tag{5.19}$$

Model (5.19) is similar to (2.2). Like (2.2), it involves N parameters, $\theta_1, \theta_2, \ldots, \theta_N$ that can be estimated by generalized linear regression. It has the advantage of taking more careful account of the observed pattern of losses and deaths within each interval, since these affect r_i but not n_i. This made little difference in the cancer study, but might be more important if the data discretization were more drastic.

Still another discretization was used by Thompson (1977), Prentice and Gloeckler (1978), and Pierce, Stewart, and Kopecky (1979). All of these methods, including (2.2), reduce the survival distribution to N unknown parameters. Here we have considered still smaller models, such as the five-parameter cubic-linear spline, to better estimate the survival distribution. The other references concentrate on the situation where there is covariate information, such as in (5.17), and do not consider parametric survival models such as (2.8).

Remark K. Several authors have investigated semiparametric hazard estimates $\hat{h}(t)$, which are smooth functions of t but do not assume a specific parametric form such as (5.4). Tanner and Wong (1983, 1984) adapted kernel estimators for use with censored data. Anderson and Senthilselvan (1980, 1982) fit the hazard by smoothing splines. O'Sullivan (1986) pursued the spline approach in a detailed study of the method of penalized likelihood. The local-regression smoother LTSM considered in Table 3 is in the spirit of these papers. Table 3 illustrates both the promise and the possible pitfalls of semiparametric smoothing techniques. The simple parametric modeling discussed in this article is a less ambitious tool that works to best advantage when the data are sparse, as in the later months of the cancer study.

Another semiparametric approach to estimating the hazard rate assumes that $h(t)$ is a monotone function of t (see Mykytyn and Santner 1981; Padgett and Wei 1980; Tsai 1986). These papers essentially use isotonic regression to estimate the binomial parameters h_i in (2.2). Miller (1981, p. 15) warned against assumptions of a monotone hazard rate in biostatistical situations, and in fact the hazard rates for the cancer study were nonmonotone. In cases where monotonicity is a safe assumption, however, the isotonic procedures are very attractive. As Tsai (1986) pointed out, the isotonic fitting algorithm continues to have a simple form even if the data are left-truncated as well as right-censored. The same statement applies to all of the methods proposed in this article.

The class of generalized linear models (McCullagh and Nelder 1984) includes logistic regression. The methods of this article allow the generalized-linear-model technology to be applied to survival analysis. McCullagh and Nelder (1984, chap. 9) presented a different way of attacking the same problem.

[*Received February 1987. Revised September 1987.*]

REFERENCES

Anderson, J. A., and Senthilselvan, A. (1980), "Smooth Estimates for the Hazard Function," *Journal of the Royal Statistical Society,* Ser. B, 42, 322–327.

——— (1982), "A Two-step Regression Model for Hazard Functions," *Applied Statistics,* 31, 44–51.

Clayton, D. (1983), "Fitting a General Family of Failure Time Distributions Using GLIM," *Applied Statistics,* 32, 102–109.

Cox, D. R. (1970), *Analysis of Binary Data,* London: Methuen.

——— (1975), "Partial Likelihood," *Biometrika,* 62, 269–278.

Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data,* London: Chapman & Hall.

Efron, B. (1977), "The Efficiency of Cox's Likelihood Function for Censored Data," *Journal of the American Statistical Association,* 72, 557–565.

——— (1986), "Logistic Regression, Survival Analysis, and the Kaplan–Meier Curve," Technical Report 266, Stanford University, Dept. of Statistics, and Technical Report 115, Stanford University, Div. of Biostatistics.

Friedman, J., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association,* 76, 817–823.

Hastie, T. J., and Tibshirani, R. J. (1986), "Generalized Additive Models," *Statistical Science,* 1, 297–318.

Holford, T. (1976), "Life Tables With Concomitant Information," *Biometrics,* 32, 587–597.

Johnson, N., and Elandt-Johnson, R. (1980), *Survival Models and Data Analysis,* New York: John Wiley.

Johnson, N., and Kotz, S. (1970), *Continuous Univariate Distributions* (Vol. 2), New York: John Wiley.

Laird, N., and Oliver, O. (1981), "Covariance Analysis of Censored Survival Data Using Log-Linear Analysis Techniques," *Journal of the American Statistical Association,* 76, 231–240.

McCullagh, P., and Nelder, J. (1984), *Generalized Linear Models,* London: Chapman & Hall.

Miller, R. G., Jr. (1981), *Survival Analysis,* New York: John Wiley.

——— (1983), "What Price Kaplan–Meier?" *Biometrics,* 39, 1077–1081.

Mykytyn, S. W., and Santner, T. J. (1981), "Maximum Likelihood Estimation of the Survival Function Based on Censored Data Under Hazard Rate Assumptions," *Communications in Statistics, Part A—Theory and Methods,* 10, 1369–1387.

O'Sullivan, F. (1986), "Estimation of Densities and Hazards by the Method of Penalized Likelihood," Technical Report 68, University of California, Berkeley, Dept. of Statistics.

Padgett, W. J., and Wei, L. J. (1980), "Maximum Likelihood Estimation of a Distribution Function With Increasing Failure Rate Based on Censored Observations," *Biometrika,* 67, 470–474.

Pierce, D., Stewart, W., and Kopecky, K. (1979), "Distribution-Free Regression Analysis of Grouped Survival Data," *Biometrics,* 35, 785–793.

Prentice, R., and Gloeckler, L. (1978), "Regression Analysis of Grouped Survival Data With Application to Breast Cancer Data," *Biometrics,* 34, 57–68.

Prentice, R., and Kalbfleisch, J. (1980), *The Statistical Analysis of Failure Time Data,* New York: John Wiley.

Rao, C. R. (1973), *Linear Statistical Inference and Its Applications,* New York: John Wiley.

Tanner, M. A., and Wong, W. H. (1983), "The Estimation of the Hazard

Function From Randomly Censored Data by the Kernel Method," *The Annals of Statistics,* 11, 989–993.

——— (1984), "Data-Based Nonparametric Estimation of the Hazard Function With Applications to Model Diagnostics and Exploratory Analysis," *Journal of the American Statistical Association,* 79, 174–182.

Thompson, W. (1977), "On the Treatment of Grouped Observations in Life Studies," *Biometrics,* 35, 463–470.

Tsai, W.-Y. (1986), "Estimation of the Survival Function With Increasing Failure Rate Based on Left Truncated and Right Censored Data," unpublished manuscript, New York: Brookhaven National Laboratory, Applied Mathematics Dept.