Published in final edited form as:

Lifetime Data Anal. 2016 January ; 22(1): 38–62. doi:10.1007/s10985-015-9326-z.

Proportional exponentiated link transformed hazards (ELTH) models for discrete time survival data with application

Hee-Koung Joeng¹, Ming-Hui Chen¹, and Sangwook Kang²

¹ Department of Statistics, University of Connecticut, Storrs, CT 06269, USA

² Department of Applied Statistics, Yonsei University, Seoul 120-749, South Korea

Abstract

Discrete survival data are routinely encountered in many fields of study including behavior science, economics, epidemiology, medicine, and social science. In this paper, we develop a class of proportional exponentiated link transformed hazards (ELTH) models. We carry out a detailed examination of the role of links in fitting discrete survival data and estimating regression coefficients. Several interesting results are established regarding the choice of links and baseline hazards. We also characterize the conditions for improper survival functions and the conditions for existence of the maximum likelihood estimates under the proposed ELTH models. An extensive simulation study is conducted to examine the empirical performance of the parameter estimates under the Cox proportional hazards model by treating discrete survival times as continuous survival times, and the model comparison criteria, AIC and BIC, in determining links and baseline hazards. A SEER breast cancer dataset is analyzed in details to further demonstrate the proposed methodology.

Keywords

C-log-log link; Cure rate models; Maximum likelihood estimate (MLE); Logit; Probit; SEER breast cancer data; t -link

1 Introduction

Discrete time survival data are common in social science, behavior science, economics, and biomedical science. Discrete time survival data analysis can be found in many applications in the literature, such as a study of family reunification using the foster care data in Wulczyn et al. (2011), the etiology of early-onset smoking behavior using the longitudinal data in Fairchild et al. (2013), the bankruptcy prediction using the financial data in Nam et al. (2008), the host selection of spring tiphia and summer tiphia using the discrete survival data

Ming-Hui Chen ming-hui.chen@uconn.edu.

Electronic supplementary material The online version of this article (doi:10.1007/s10985-015-9326-z) contains supplementary material, which is available to authorized users.

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

of oriental beetles grubs in Obeysekara (2013), and an analysis of the discrete survival data from the study of Viral Resistance to Antiviral Therapy of Chronic Hepatitis C (Virahep-C) in Adeniji et al. (2014). To motivate the proposed methodology, we consider a subset of the Surveillance, Epidemiology, and End Results (SEER) data from the National Cancer Institute (NCI) with 2096 female subjects who were at least 20 years old at diagnosis and were diagnosed with regional extension, grade III or IV, and stage III breast cancer between 1990 and 2003. In SEER data, only the month and year but not the day of death or last follow-up or diagnosis are available due to patient confidentiality consideration. Thus, it is quite natural to consider these event times as discrete survival times. In the aforementioned applications, there are two types of discrete survival data. The first type is that the survival time is originally continuous but recorded in a discrete version by grouping or rounding into a discrete time. We call this type as “derived discrete”. The survival time in the SEER data is an example of this type since the survival time in the unit of months is a rounded number due to patient confidentiality consideration. The second type is “intrinsically discrete”. An example of the intrinsically discrete time survival data is given in Adeniji et al. (2014). In Adeniji et al. (2014), the viral levels ($w(t)$) of Hepatitis C virus (HCV) were measured at eight discrete time points, $t_1, \dots, t_8$, over 48 weeks and an event of interest occurred only when the viral level was below a detection limit at the first time over these eight visits. In this case, the information about the underlying viral process $w(t)$ was observed only at $t_1, \dots, t_8$, but no information about $w(t)$ was available between any two time points t_j and t_{j+1} . This type of discrete survival time arises when it is not possible to monitor the underlying viral process $w(t)$ at every time t in clinical practice.

Allison (1982) discussed a logistic regression model for discrete time hazards. Kalbfleisch and Prentice (1973) developed a discrete hazard model for grouped data. Other statistical methods for modeling discrete time survival data have been developed in the literature, including Prentice and Gloeckler (1978), Stewart and Pierce (1982), Efron (1988), Singer and Willet (1993), Biggeri et al. (2001), Grilli (2005), Muthen and Masyn (2005), Manda and Meyer (2005), Brown et al. (2009), and Nguyen and Gillen (2012). Steele (2003) developed the discrete time multi-level mixture model with long term survivors, which allows random effects for unobserved individual heterogeneity in the hazard of event occurrence. Zhao and Zhou (2008) proposed a discrete time proportional hazards model with long term survivors as an extension of the mixed discrete and continuous Cox's regression model in Prentice and Kalbfleisch (2003) by allowing improper baseline survival function.

In this paper, we develop a new class of proportional exponentiated link transformed hazards (ELTH) models for discrete time survival data. The ELTH model is attractive since it mimics the Cox proportional hazards model for continuous failure times and the class of ELTH models are flexible and rich, which include the logistic regression model (Allison 1982), the continuous-time proportional hazards generated model of Kalbfleisch and Prentice (1973), and the cure rate model of Zhao and Zhou (2008) as special cases. The proposed approach is quite different than those developed in Cheng et al. (1995), Chen et al. (2002), and Banerjee et al. (2007). Specifically, in the class of ELTH models, we make the transformation directly on the hazard function while Cheng et al. (1995) and Banerjee et al. (2007) considered a class of transformations on the survival function and Chen et al. (2002)

developed the linear transformation models on the failure time. We note here that the transformation approach on the failure time considered in Chen et al. (2002) is not directly applicable to discrete survival data since the distribution of a discrete failure time is unchanged after a one-to-one monotone transformation. As elaborated in Sect. 2, discrete time survival data can be viewed as binary response data. Thus, the methods for fitting binary response data in the existing literature may be directly used for discrete survival data. However, in many applications, especially those that arise in biomedical studies, the hazards are typically small. Thus, the choice of links for discrete time survival data is quite different than the one for general binary response data. The ELTH model with a symmetric link may fit the discrete time survival data as equally well as the one with a skewed link. In addition, the ELTH models with very different links may lead to similar estimates of the regression coefficients. In this regard, we establish an interesting result to characterize the role of links in fitting discrete time survival data. Another contribution of the paper is to characterize the conditions for improper survival functions under the ELTH models. These conditions are easy to check and help the data analyst to determine and compute subject-specific cure rates. We further examine the conditions for existence of the maximum likelihood estimates. The establishment of these conditions is of practical significance in determining whether the proposed ELTH models are identifiable in fitting discrete time survival data in real applications.

The rest of the paper is organized as follows. In Sect. 2, we provide a detailed development of the proposed ELTH models, examine various properties of the proposed models including the choice of links and the conditions for improper survival functions, and present the likelihood function and the conditions for existence of the maximum likelihood estimates. An extensive simulation study is carried out in Sect. 3. A detailed analysis of the SEER breast cancer data is given in Sect. 4. We conclude the paper with some discussion and future research in Sect. 5. Some useful remarks, the proofs of theorems, and additional tables are given in the online supplementary materials (Appendix A–Appendix C in Electronic Supplementary Material).

2 The methods

2.1 Preliminary

Let T be a discrete random variable taking only positive values $0 < t_1 < t_2 < \dots$. Then the discrete time hazard function at t_k is defined as

$$h(k) = P(T = t_k | T \geq t_k). \quad (2.1)$$

After some algebra, the probability of $T = t_k$ can be expressed as

$$P(T = t_k) = h(k) \prod_{j=1}^{k-1} \{1 - h(j)\}, \quad k=1, 2, \dots, \quad (2.2)$$

where $\prod_{j=1}^0 \{1 - h(j)\} = 1$. The corresponding survival function at time t_k is given by

$$S(k) = P(T > t_k) = \prod_{i=1}^k \{1 - h(i)\} = \exp \left[\sum_{i=1}^k \log \{1 - h(i)\} \right] \quad (2.3)$$

for $k = 1, 2, \dots$. For an arbitrary time t , the discrete time survival function can be written as follows

$$\prod_{i:t_i \leq t} \{1 - h(i)\} \equiv S(k_t),$$

where $k_t = \max\{i : t_i \leq t\}$. Let $\mathbf{x}$ denote a vector of baseline covariates and also let $\boldsymbol{\beta}$ be a vector of the corresponding regression coefficients. To develop a regression model, Cox (1972) proposed the following discrete time hazard regression model:

$$\frac{h(k|\mathbf{x}, \boldsymbol{\beta})}{1 - h(k|\mathbf{x}, \boldsymbol{\beta})} = \frac{h_0(k)}{1 - h_0(k)} \exp(\mathbf{x}'\boldsymbol{\beta}), \quad (2.4)$$

where $h_0(k)$ is the baseline discrete hazard function at time t_k . Taking the logarithm of both sides of (2.4) gives

$$\text{logit}(h(k|\mathbf{x}, \boldsymbol{\beta})) \equiv \log \left\{ \frac{h(k|\mathbf{x}, \boldsymbol{\beta})}{1 - h(k|\mathbf{x}, \boldsymbol{\beta})} \right\} = \log \left\{ \frac{h_0(k)}{1 - h_0(k)} \right\} + \mathbf{x}'\boldsymbol{\beta}. \quad (2.5)$$

Letting $\alpha_k = \log \left\{ \frac{h_0(k)}{1 - h_0(k)} \right\}$, we have

$$h(k|\mathbf{x}, \boldsymbol{\beta}) = \frac{\exp(\alpha_k + \mathbf{x}'\boldsymbol{\beta})}{1 + \exp(\alpha_k + \mathbf{x}'\boldsymbol{\beta})}$$

for $k = 1, 2, \dots$. The logistic regression model in (2.5) is one of the most popular models for discrete survival time data (e.g., see Allison 1982).

Kalbfleisch and Prentice (1973) developed the discrete model for the grouped data from the Cox's model with the hazard function given by

$$h(k|\mathbf{x}, \boldsymbol{\beta}) = 1 - [1 - h_0(k)]^{\exp(\mathbf{x}'\boldsymbol{\beta})}, \quad (2.6)$$

where the survival time in the interval $[a_{k-1}, a_k]$ is recorded as t_k , $a_0 = 0 < a_1 < \dots$, and $h_0(k)$ is the baseline discrete hazard function. From (2.6), we have

$$\log[-\log\{1 - h(k|\mathbf{x}, \boldsymbol{\beta})\}] = \log[-\log\{1 - h_0(k)\}] + \mathbf{x}'\boldsymbol{\beta} \quad (2.7)$$

and

$$h(k|\mathbf{x}, \boldsymbol{\beta}) = 1 - \exp \left[-\exp(\alpha_k + \mathbf{x}'\boldsymbol{\beta}) \right], \quad (2.8)$$

where $\alpha_k = \log[-\log\{1 - h_0(k)\}]$ for $k = 1, 2, \dots$. As discussed in Allison (1982), the discrete time hazard function is (2.8) if the data are generated by the continuous time proportional hazards model. Another discrete hazard regression model proposed by Zhao and Zhou (2008) assumes

$$h(k|\mathbf{x}, \boldsymbol{\beta}) = 1 - \left[\frac{1 - \pi + \pi \prod_{j=1}^k (1 - h_0^Z(j))}{1 - \pi + \pi \prod_{j=1}^{k-1} (1 - h_0^Z(j))} \right]^{\exp(\mathbf{x}'\boldsymbol{\beta})}, \quad (2.9)$$

where $0 < h_0^Z(k) < 1$ for $k = 1, 2, \dots$, $0 < \pi < 1$, and $\prod_{j=1}^0 \{1 - h_0^Z(j)\} = 1$. We note that (2.9) is more general than (2.6) since (2.9) reduces to (2.6) where $\pi = 1$ and $h_0^Z(k) = h_0(k)$.

Let $S_0^Z(k) = \prod_{j=1}^k (1 - h_0^Z(j))$, which defines a discrete time survival function. Then, Zhao and Zhou (2008) assumed that $S_0^Z(k)$ is a proper survival function, that is, $S_0^Z(\infty) = \lim_{k \rightarrow \infty} S_0^Z(k) = 0$, which implies

$$\prod_{j=1}^{\infty} (1 - h_0^Z(j)) = 0. \quad (2.10)$$

In general, any hazard function that satisfies (2.10) defines a proper survival function. Let $S(k|\mathbf{x}, \boldsymbol{\beta})$ denote the survival function corresponding to (2.9). Then, using (2.1) and (2.3), we have

$$h(k|\mathbf{x}, \boldsymbol{\beta}) = 1 - \frac{S(k|\mathbf{x}, \boldsymbol{\beta})}{S(k-1|\mathbf{x}, \boldsymbol{\beta})}. \quad (2.11)$$

Combining (2.9) and (2.11) leads to

$$S(k|\mathbf{x}, \boldsymbol{\beta}) = \left[1 - \pi + \pi \prod_{j=1}^k \{1 - h_0^Z(j)\} \right]^{\exp(\mathbf{x}'\boldsymbol{\beta})} = \left[1 - \pi + \pi S_0^Z(k) \right]^{\exp(\mathbf{x}'\boldsymbol{\beta})}. \quad (2.12)$$

From (2.12), it is easy to see that $S(\infty|\mathbf{x}, \boldsymbol{\beta}) = (1 - \pi)^{\exp(\mathbf{x}'\boldsymbol{\beta})}$ since $S_0^Z(\infty) = 0$. Thus, if $\pi < 1$, $S(k|\mathbf{x}, \boldsymbol{\beta})$ is not a proper survival function. Now, it is also clear that $1 - \pi$ represents the cured fraction of the baseline population.

2.2 The proposed models

From (2.1), we see that the discrete time hazard rate function $h(k)$ is the conditional probability of an event of interest occurring at time t_k , given that the event has not occurred yet prior to t_k and, hence, $0 < h(k) < 1$. When there are subject-dependent covariates ($\mathbf{x}$), one of the key issues is how to model $h(k)$, i.e., how to link $h(k)$ to $\mathbf{x}$, for discrete survival data. By extending (2.5) and (2.7), we propose the following additive model for the hazard function:

$$F^{-1}(h(k|\mathbf{x}, \boldsymbol{\beta})) = \alpha_k + \mathbf{x}'\boldsymbol{\beta}, \quad (2.13)$$

where F is a continuous cumulative distribution function (cdf) and F^{-1} is the inverse function of F . Throughout the paper, we assume $0 < F(u) < 1$ for $-\infty < u < \infty$. In (2.13),

F^{-1} is called a link function. When $F^{-1}(u) = \text{logit}(u) = \log\left(\frac{u}{1-u}\right)$ (logit), (2.13) reduces to (2.5) with $\alpha_k = \text{logit}(h_0(k))$ and similarly, when $F^{-1}(u) = \log\{-\log(1-u)\}$ (C-log-log), (2.13) reduces to (2.7) with $\alpha_k = \log[-\log\{1-h_0(k)\}]$, where $h_0(k)$ is the baseline hazard function. Setting $\alpha_k = F^{-1}(h_0(k))$, we can rewrite (2.13) as follows:

$$\exp\{F^{-1}(h(k|\mathbf{x}, \boldsymbol{\beta}))\} = \exp\{F^{-1}(h_0(k))\} \exp(\mathbf{x}'\boldsymbol{\beta}). \quad (2.14)$$

The model in the form of (2.14) is attractive since it mimics the Cox proportional hazards model for continuous failure times. Thus, the proposed model can be viewed as the proportional exponentiated link transformed hazards (ELTH) model and α_k is called the link transformed baseline hazard function. If we take $F^{-1}(u) = \log\{-\log(1-u)\}$ and let

$$h_0(k) = \frac{\pi h_0^Z(k) \prod_{j=1}^{k-1} \{1 - h_0^Z(j)\}}{1 - \pi + \pi \prod_{j=1}^{k-1} \{1 - h_0^Z(j)\}}, \quad (2.15)$$

where $h_0^Z(k)$ is defined in (2.9), the model in (2.13) or (2.14) reduces to (2.9). Therefore, the cure rate model of Zhao and Zhou (2008) is a special case of the ELTH model.

2.3 The choices of links

From (2.2) and (2.3), we see that events $\{T = t_k\}$ and $\{T > t_k\}$ are equivalent to events $\{Y_1 = 0, \dots, Y_{k-1} = 0, Y_k = 1\}$ and $\{Y_1 = 0, \dots, Y_{k-1} = 0, Y_k = 0\}$, respectively, where $Y_1, \dots, Y_k$ are k binary random variables taking values of 0 or 1 such that the Y_ℓ 's are independent and $Y_\ell \sim \text{Bernoulli}(h(\ell))$ for $\ell = 1, \dots, k$. Thus, discrete survival time data can be viewed as binary response data. For binary response data, as discussed in Stukel (1988), Czado and Santner (1992), Chen et al. (1999), and Kim et al. (2008), the misspecification of the link may yield substantial bias in the mean response estimates and a skewed link may fit the data much better than a symmetric link or vice versa. However, for discrete survival data, the hazards, $h(k)$'s, are typically small. As shown in Fig. 5, all of the naïve estimates of the hazards are less than 0.02 for the SEER breast cancer data. When the hazards are less than 0.5, the half of the link corresponding to the portion of the cumulative distribution function (cdf) F with $F = 0.5$ does not play any role in fitting discrete survival data. We now state a useful proposition.

Empirical Observation—Let F_1 and F_2 be two cdf's. Let $(\alpha_k^{(1)}, \boldsymbol{\beta}^{(1)})$ and $(\alpha_k^{(2)}, \boldsymbol{\beta}^{(2)})$ be the respective parameters corresponding to F_1 and F_2 in (2.13). If there exist constants $c_1 > 0$ and c_2 such that (i) $\lim_{u \rightarrow -\infty} [F_1(c_1 u + c_2) / F_2(u)] = 1$ or (ii) $\lim_{u \rightarrow -\infty} [F_1(u) / F_2(c_1 u + c_2)] = 1$, then we have $\alpha_k^{(1)} \approx c_1 \alpha_k^{(2)} + c_2$ and $\boldsymbol{\beta}^{(1)} \approx c_1 \boldsymbol{\beta}^{(2)}$ for (i) and

$\alpha_k^{(2)} \approx c_1 \alpha_k^{(1)} + c_2$ and $\beta^{(2)} \approx c_1 \beta^{(1)}$ for (ii), where “ $\approx$ ” denotes “approximately equals”, when the hazards are small.

We note that when the hazards are sufficiently small, $F_1(c_1 u + c_2)/F_2(u)$ (or $F_1(u)/F_2(c_1 u + c_2)$) will be close to 1. For example, when $F_1^{-1}(u) = \log\{-\log(1-u)\}$ and

$F_2^{-1}(u) = \log\left(\frac{u}{1-u}\right)$, we obtain that $F_1(u) = 0.08$, $F_2(u) = 0.08$, and $|F_1(u)/F_2(u) - 1| = 0.04$ for $u = -2.5$; $F_1(u) = 0.05$, $F_2(u) = 0.05$, and $|F_1(u)/F_2(u) - 1| = 0.025$ for $u = -3$; and $F_1(u) = 0.02$, $F_2(u) = 0.02$, and $|F_1(u)/F_2(u) - 1| = 0.01$ for $u = -4$. Based on our empirical

experience, the estimates under the ELTH model with the logit link are practically the same as those under the ELTH model under the C-log-log link when the hazards are less than 0.08. Thus, the hazards would be considered to be small when they are about 0.08 or less.

The practical implication of our Empirical Observation is that the ELTH model with a symmetric link may fit discrete survival data as equally well as the ELTH model with an asymmetric link when the hazards are small. These properties will be examined in details in the simulation study in Sect. 3 and a real data analysis in Sect. 4.

Figure 1 shows the cdf plots for the probit, logit, C-log-log, t_4 , and standard Laplace links, where F is the standard normal cdf Φ under the probit link, F is the standard t -distribution function with 4° of freedom under the t_4 link, and the standard Laplace cdf is given by

$F(u) = \frac{1}{2} \exp(u)$ if $u < 0$ and $1 - \frac{1}{2} \exp(-u)$ if $u \geq 0$. Among these five links, the C-log-log link is skewed while other four are symmetric. If we compare the entire curves shown in Plot (a) of Fig. 1, all five cdf's are different. If we compare the curves shown in Plot (b) of Fig. 1, we observe that (i) the C-log-log and logit links are almost indistinguishable; (ii) the t_4 (probit) cdf approaches to 0 at a much slower (faster) rate than the cdf's under other four links; and (iii) the standard Laplace cdf approaches to 0 at a faster rate than both the C-log-log and logit cdf's but a slower rate than the probit cdf. Let $F_1(u) = \exp(u)/(1 + \exp(u))$, $F_2(u) = 1 - \exp\{-\exp(u)\}$, and $F_3(u) = \frac{1}{2} \exp(u)$ for $u < 0$ be the cdf's corresponding to the C-log-log, logit, and standard Laplace links. Then, we can show that $\lim_{u \rightarrow -\infty} [F_1(u)/F_2(u)] = 1$ and $\lim_{u \rightarrow -\infty} [F_1(u)/F_3(u + \log 2)] = 1$. Although the C-log-log link is different than the logit link as shown in Plot (a) of Fig. 1 as well as the standard Laplace link as shown in Plot (b) of Fig. 1, these three links will yield similar estimates of the regression coefficients (β) and a similar fit of the discrete survival data when the hazards are small according to our Empirical Observation. Thus, the choice of links in fitting discrete survival data is quite different than the one in fitting general binary response data.

2.4 The cure rates under the ELTH models

Using (2.10) and (2.15), it is easy to see that

$$\lim_{k \rightarrow \infty} h_0(k) = 0 \quad \text{and} \quad \lim_{k \rightarrow \infty} \alpha_k = \lim_{k \rightarrow \infty} \log[-\log\{1 - h_0(k)\}] = -\infty \quad (2.16)$$

if $\pi < 1$. Thus, the cure rate model implies (2.16).

Next, we examine the conditions on F and $h_0(k)$ under which the ELTH model induces a cure rate model. We first state two assumptions on α_k and F as follows.

Assumption 1— $\lim_{k \rightarrow \infty} \alpha_k = -\infty$ and for a given c , there exist $k_0 > 0$ and $d > 0$ such that $\alpha_k + c < 0$ and

$$F(\alpha_k + c) \leq d \exp(\alpha_k + c) \quad \text{for all } k \geq k_0.$$

Assumption 2— $\lim_{k \rightarrow \infty} \alpha_k = -\infty$ and for a given $-\infty < c < \infty$, there exist $k_0 > 0$, $d > 0$, and $r > 0$ such that $\alpha_k + c < 0$ and

$$F(\alpha_k + c) \leq \frac{d}{(-\alpha_k - c)^r} \quad \text{for all } k \geq k_0.$$

It is easy to see that if F satisfies Assumption 1, F automatically satisfies Assumption 2. Thus, the conditions in Assumption 1 are stronger than those in Assumption 2. Using (2.3) and (2.13), we have

$$S(k|\mathbf{x}, \boldsymbol{\beta}) = \exp \left[\sum_{i=1}^k \log \left\{ 1 - F(\alpha_i + \mathbf{x}'\boldsymbol{\beta}) \right\} \right]. \quad (2.17)$$

If $S(k|\mathbf{x}, \boldsymbol{\beta})$ is not proper, i.e., $S(\infty|\mathbf{x}, \boldsymbol{\beta}) > 0$, then we must have

$$\lim_{k \rightarrow \infty} \log \left\{ 1 - F(\alpha_k + \mathbf{x}'\boldsymbol{\beta}) \right\} = 0. \quad (2.18)$$

When $F'(u) = \frac{dF(u)}{du} > 0$, (2.18) implies $\lim_{k \rightarrow \infty} \alpha_k = -\infty$. Consequently, we have $\lim_{k \rightarrow \infty} h_0(k) = 0$ since $h_0(k) = F(\alpha_k)$. Thus, the condition, $\lim_{k \rightarrow \infty} \alpha_k = -\infty$, is necessary for an improper survival function. Now, we establish the sufficient conditions for an improper survival function in the following theorem.

Theorem 1—The survival function $S(k|\mathbf{x}, \boldsymbol{\beta})$ is improper, i.e., $S(\infty|\mathbf{x}, \boldsymbol{\beta}) > 0$ if (i) α_k and F satisfy Assumption 1 and

$$\sum_{k=1}^{\infty} \exp(\alpha_k) < \infty; \quad (2.19)$$

or (ii) α_k and F satisfy Assumption 2 with $c = \mathbf{x}'\boldsymbol{\beta}$ and

$$\sum_{k=k_0}^{\infty} \frac{1}{(-\alpha_k - c)^r} < \infty. \quad (2.20)$$

The proof of Theorem 1 is given in the Appendix in Electronic Supplementary Material. For certain link functions, (2.19) becomes the necessary and sufficient condition for an improper survival function. We formally state this result in the following theorem.

Theorem 2—The survival function $S(k|\mathbf{x}, \boldsymbol{\beta})$ is improper if and only if (i) (2.19) holds when F is the cdf corresponding to the C-log-log link or the logit link and (ii) (2.20) with $r = v$ holds when F is the cdf of a t-distribution with v degrees of freedom.

The proof of Theorem 2 is given in the Appendix in Electronic Supplementary Material. In the next theorem, we present a sufficient condition for a proper survival function.

Theorem 3—The survival function $S(k|\mathbf{x}, \boldsymbol{\beta})$ is proper, i.e., $S(\infty|\mathbf{x}, \boldsymbol{\beta}) = 0$ if

$$\overline{\lim}_{k \rightarrow \infty} \alpha_k > -\infty.$$

The proof of this theorem is straightforward and thus the detail of the proof is omitted for brevity. As discussed in Remark A.3 of Appendix A in Electronic Supplementary Material, there is a connection between the cure rate model (2.9) and the ELTH model. This connection indicates that the cure rate model can be a direct consequence of the ELTH model by properly modeling the baseline hazard $h_0(k)$ or the link transformed baseline hazard α_k , which will be discussed in detail in the next subsection.

2.5 Modeling the baseline hazard functions

The simplest model for α_k is the linear model (e.g., see Allison 1982) given by

$$\alpha_k = \psi_0 + \psi_1 t_k, \quad (2.21)$$

where t_k is the time at which the survival function $S(k|\mathbf{x}, \boldsymbol{\beta})$ is defined. The following theorem characterizes the conditions for proper or improper survival functions under the linear model (2.21).

Theorem 4—Let $\mathcal{N}_j = \{k: t_k \in (j-1, j]\}$ for $j = 1, 2, \dots$. Suppose $t_k \rightarrow \infty$ as $k \rightarrow \infty$ and $N_{\max} = \sup_{j \geq 1} |\mathcal{N}_j| < \infty$, where $|\mathcal{N}_j|$ denotes the cardinal number of $\mathcal{N}_j$. Then, we have (i) $S(\infty|\mathbf{x}, \boldsymbol{\beta}) > 0$ for F satisfying Assumption 1 or Assumption 2 with $r > 1$ if $\psi_1 < 0$; and (ii) $S(\infty|\mathbf{x}, \boldsymbol{\beta}) = 0$ for any F if $\psi_1 = 0$.

By extending the idea in Efron (1988) and (2.21), we further consider a piecewise baseline hazard model for α_k . We first construct a finite partition of the interval $[t_1, \infty)$, $s_0 = t_1 < s_1 < s_2 < \dots < s_{J-1} < s_J = \infty$. Thus, we have J intervals $[t_1, s_1)$, $[s_1, s_2)$, $\dots$, $[s_{J-1}, s_J)$. We then assume piecewise constants for α_k for the first $J-1$ intervals and a linear function for α_k in the last interval, that is,

$$\alpha_k = \begin{cases} \psi_0 + \psi_1 t_k & \text{if } J=1, \\ \psi_0 + \sum_{j=1}^{J-2} \psi_j I\{s_j \leq t_k < s_{j+1}\} + \psi_{J-1} (t_k - s_{J-1})_+ & \text{if } J>1, \end{cases} \quad (2.22)$$

where $I\{s_j \leq t_k < s_{j+1}\}$ is the indicator function, which takes a value of 1 if $s_j \leq t_k < s_{j+1}$ and 0 otherwise, and $(t_k - s_{J-1})_+ = t_k - s_{J-1}$ if $t_k > s_{J-1}$ and 0 otherwise. From (2.22), we see that for $J > 1$, $\alpha_k = \psi_0$ for $t_k < s_1$ and $\psi_0 + \psi_j$ for $s_j \leq t_k < s_{j+1}$ for $j = 1, \dots, J-2$ (i.e., piecewise constants) and $\alpha_k = \psi_0 + \psi_{J-1}(t_k - s_{J-1})$ (linear) for $t_k \geq s_{J-1}$. When $J = 1$, (2.22) reduces (2.21). The conditions for proper or improper survival functions under the piecewise baseline hazard model (2.22) are similar to those given in Theorem 4.

The form for α_k in (2.22) is quite general and flexible since it can capture any shapes of the unknown true baseline hazard $h_0(k)$ in (2.14) by increasing J . For a given dataset, the ELTH model with α_k in (2.22) may fit the data poorly when J is too small and the fit of the ELTH model improves when J is large. However, the number of unknown parameters ψ_j 's becomes large and the model may become too complex and even unidentifiable when J is unnecessarily large. Thus, there is a trade-off between the goodness-of-fit and the model complexity in the choice of J . To address these issues, we characterize the conditions for the existence of the maximum likelihood estimates of the ψ_j 's as well as β in Theorem 5 and propose the Akaike information criterion (AIC) and the Bayesian Information Criterion (BIC) to guide the choice of J in Sect. 2.8 as well as in the simulation study and the analysis of the breast cancer data.

2.6 The likelihood function

Suppose there are n subjects. For the i th subject, let t_{ki} denote the discrete failure time, which may be right-censored, and also let δ_i denote the censoring indicator such that $\delta_i = 1$ if t_{ki} is the failure time and 0 otherwise. We further let $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ denote a p -dimensional vector of covariates. Write $D = \{(t_{ki}, \delta_i, \mathbf{x}_i), i = 1, \dots, n\}$, which is the observed data. Then, the observed data likelihood function under the ELTH model can be written as

$$\begin{aligned} L(\beta, \mathbf{h}_0 | D) &= \prod_{i=1}^n [Pr(T_i = t_{ki} | \mathbf{x}_i, \beta, \mathbf{h}_0)]^{\delta_i} [Pr(T_i > t_{ki} | \mathbf{x}_i, \beta, \mathbf{h}_0)]^{1-\delta_i} \\ &= \prod_{i=1}^n \left[\frac{h(k_i | \mathbf{x}_i, \beta)}{1 - h(k_i | \mathbf{x}_i, \beta)} \right]^{\delta_i} S(k_i | \mathbf{x}_i, \beta), \end{aligned} \quad (2.23)$$

where β is an p -dimensional vector of regression coefficients corresponding to $\mathbf{x}_i$, h_0 is the baseline hazard function, and $h(k_i | \mathbf{x}_i, \beta)$ and $S(k_i | \mathbf{x}_i, \beta)$ are defined by (2.13) and (2.17), respectively. After some algebra, (2.23) can be rewritten as

$$\begin{aligned} L(\beta, \mathbf{h}_0 | D) &= \prod_{i=1}^n \left\{ \frac{F(\alpha_{k_i} + \mathbf{x}_i' \beta)}{1 - F(\alpha_{k_i} + \mathbf{x}_i' \beta)} \right\}^{\delta_i} \prod_{j=1}^{k_i} \{1 - F(\alpha_j + \mathbf{x}_i' \beta)\} \\ &= \prod_{i=1}^n \left[\left\{ F(\alpha_{k_i} + \mathbf{x}_i' \beta) \right\}^{\delta_i} \left\{ 1 - F(\alpha_{k_i} + \mathbf{x}_i' \beta) \right\}^{1-\delta_i} \prod_{j=1}^{k_i-1} \{1 - F(\alpha_j + \mathbf{x}_i' \beta)\} \right]. \end{aligned}$$

2.7 The existence of the MLE

Now, we characterize the conditions for the existence of the maximum likelihood estimate (MLE). For ease of exposition, we consider the piecewise baseline hazard model (2.22) for

the α_k . Let $\mathbf{z}_{im} = (1, t_m, \mathbf{x}_i')$ for $J = 1$, $\mathbf{z}_{im} = (1, (t_m - s_1)_+, \mathbf{x}_i')$ for $J = 2$, and

$\mathbf{z}_{im} = (1, I\{s_1 \leq t_m < s_2\}, \dots, I\{s_{J-2} \leq t_m < s_{J-1}\}, (t_m - s_{J-1})_+, \mathbf{x}_i')$ for $J > 2$, $m = 1,$

$2, \dots, k_i$ and $i = 1, 2, \dots, n$. Also let $\mathbf{z}_{im}^* = \mathbf{z}_{im}$ for $m = 1, 2, \dots, k_i - 1$ and $\mathbf{z}_{ik_i}^* = (1 - 2\delta_i) \mathbf{z}_{ik_i}$

for $i = 1, 2, \dots, n$ and $N = \sum_{i=1}^n k_i$. Define $X = (\mathbf{z}_{im}^*, m=1, 2, \dots, k_i, i=1, 2, \dots, n)'$, which is a $N \times (2 + p)$ matrix for $J = 1$ or a $N \times (J + p)$ matrix for $J > 1$. We are led to the following theorem.

Theorem 5—Suppose that X is of full rank, F is continuous, and $0 < F < 1$. Then the maximum likelihood estimates (MLE) of $\boldsymbol{\psi} = (\psi_0, \dots, \psi_{J-1})'$ and $\boldsymbol{\beta}$ under the ELTH model exist if and only if there exists a N -dimensional positive vector $\mathbf{a} = (a_1, \dots, a_N)'$, i.e., $a_i > 0$ for $i = 1, \dots, N$, such that

$$X' \mathbf{a} = 0.$$

The proof of Theorem 5 directly follows from Theorem 3.1 in Chen and Shao (2001). We now present a simply example to illustrate the results established in Theorem 5.

2.7.1 An illustrative example: We consider two small discrete survival datasets (denoted by D_1 and D_2) with two covariates shown in Table 1. We note that these two datasets are two subsets of the SEER breast cancer data with covariates estrogen receptor status (ER) and treatment indicators of surgery (Surg). Due to the small sample size, we assume the piecewise baseline hazard model (2.22) for the α_k with $J = 1$. For data D_1 , $k_1 = 1, k_2 = 2, k_3 = 1, p = 2, z_{11}^* = -(1, 1, 0, 0)'$, $z_{21}^* = (1, 1, 0, 1)'$, $z_{22}^* = -(1, 2, 0, 1)'$, and $z_{31}^* = -(1, 1, 1, 1)'$.

$$X = \begin{pmatrix} -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ -1 & -2 & 0 & -1 \\ -1 & -1 & -1 & -1 \end{pmatrix}$$

Thus, $N = 4$, and X is a 4×4 matrix given by $X = \begin{pmatrix} -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ -1 & -2 & 0 & -1 \\ -1 & -1 & -1 & -1 \end{pmatrix}$. Since X is of full rank, only $\mathbf{a} = (0, 0, 0, 0)'$ satisfies $X' \mathbf{a} = 0$. Therefore, the MLE of $(\psi_0, \psi_1, \beta_1, \beta_2)$ does

not exist for data D_1 . For data D_2 , $k_1 = 1, k_2 = 2, k_3 = 1, k_4 = 3, p = 2, z_{11}^* = -(1, 1, 0, 0)'$,

$z_{21}^* = (1, 1, 0, 1)'$, $z_{22}^* = -(1, 2, 0, 1)'$, $z_{31}^* = -(1, 1, 1, 1)'$, $z_{41}^* = -(1, 1, 1, 0)'$,

$z_{42}^* = -(1, 2, 1, 0)'$, and $z_{43}^* = -(1, 3, 1, 0)'$. Thus, $N = 7$, and the corresponding matrix X is

$$X' = \begin{pmatrix} -1 & 1 & -1 & -1 & 1 & 1 & -1 \\ -1 & 1 & -2 & -1 & 1 & 2 & -3 \\ 0 & 0 & 0 & -1 & 1 & 1 & -1 \\ 0 & 1 & -1 & -1 & 0 & 0 & 0 \end{pmatrix}$$

. It is easy to verify that (i) X is of full rank and (ii) a positive vector $\mathbf{a} = (3, 4, 1, 3, 1, 3, 1)'$ satisfies $X' \mathbf{a} = 0$. Thus, the MLE of $(\psi_0, \psi_1, \beta_1, \beta_2)$ exists for data D_2 .

2.8 Model comparison and assessment

Under the piecewise baseline hazard model (2.22), the likelihood function $L(\boldsymbol{\beta}, h_0|D)$ in

(2.23) is a function of $(\boldsymbol{\psi}, \boldsymbol{\beta})$. Thus, we rewrite $L(\boldsymbol{\psi}, \boldsymbol{\beta}|D) = L(\boldsymbol{\beta}, h_0|D)$. Let $(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}})$ denote the

MLE of $(\boldsymbol{\psi}, \boldsymbol{\beta})$. Then, AIC is given by $AIC = -2 \log L(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}|D) + 2 \dim(\boldsymbol{\psi}, \boldsymbol{\beta})$, and BIC is defined as $BIC = AIC + (\log n - 2) \dim(\boldsymbol{\psi}, \boldsymbol{\beta})$. We examine the performance of AIC and BIC in determining the choice of links and the number of pieces in the piecewise baseline hazard model (2.22) in a simulation study in Sect. 3 and we also use AIC and BIC to select the best model to fit the SEER breast cancer data in Sect. 4.

3 A simulation study

We conduct two simulation studies as follows. For each simulated dataset of size n , we first generate $\mathbf{x}_i = (x_{1i}, x_{2i})'$ as $x_{1i} \sim N(0, 1)$ and $x_{2i}|x_{1i} \sim \text{Bernoulli}(q_i)$, where $q_i = \exp(0.1 - 0.1x_{1i}) / \{1 + \exp(0.1 - 0.1x_{1i})\}$. We then use the inverse cdf approach to generate the discrete survival time T_i from the ELTH model (2.14) under the logit link along with the linear baseline hazard model (2.21) for the α_k with $\boldsymbol{\psi} = (-2.7, -0.03)'$ and $\boldsymbol{\beta} = (0.5, 0.5)'$ (Simulation I) or the piecewise baseline hazard model (2.22) for the α_k with $J = 5$, $(s_1, s_2, s_3, s_4) = (3, 6, 10, 22)$, $\boldsymbol{\psi} = (-3.5, 0.2, 0.35, 0.05, -0.01)'$, and $\boldsymbol{\beta} = (0.5, 0.5)'$ (Simulation II). The censoring time C_i is generated uniformly from a discrete set $\{1, 2, \dots, 85\}$ (Simulation I) or $\{1, 2, \dots, 100\}$ (Simulation II). Finally the observed discrete survival time is $t_{ki} = \min\{T_i, C_i\}$ and the censoring indicator $\delta_i = I\{T_i \leq C_i\}$ for $i = 1, \dots, n$. In both simulation studies, we generate 1,000 simulated datasets with $n = 1,000$ or $n = 2,000$. Under the above settings, average percentage of censored observations is about 26% and all of the average hazards are less than 0.09 across 1,000 simulated datasets.

In Simulation I, for each simulated dataset, we fit the Cox proportional hazards model by treating the discrete time survival data as the continuous survival time data and compute the MLEs of $\boldsymbol{\beta}$ and the corresponding standard errors and 95% confidence intervals (CI's) under the true model as well as the Cox model; and we also fit the ELTH models under the logit (true), C-log-log, probit, and t_4 links and compute AIC and BIC for each of the four ELTH models. In Simulation II, for each simulated dataset, we fit the four ELTH models with $J = 5$ as in Simulation I and compute AIC and BIC for each model; and we also fit the ELTH models under the logit link for $J = 1$, $J = 5$ (true), $J = 10$, and $J = 15$ and again compute AIC and BIC for each J .

3.1 Results of simulation I

Table 2 shows the means of estimates (EST's), the averages of standard errors (ASE's), the simulation standard deviations of estimates (SSD's), the roots of mean squared errors (RMSE's), and the coverage probabilities (CP's) of 95% confidence intervals for β_1 and β_2 under the ELTH and Cox models. We see from Table 2 that (i) under the ELTH model with the true logit link, the estimates of $\boldsymbol{\beta}$ were very close to the true values, the SSD and ASE were almost the same for each of β_1 and β_2 , both the SSD and ASE decreased when n increases, and the CPs were close to 95%; (ii) the estimates of $\boldsymbol{\beta}$ under the ELTH model with the C-log-log link were much better than those under ELTH models with the probit and t_4 links and the CP's were 0 under the ELTH model with the probit link; and (iii) under the Cox model, the estimates were biased, and when n increases, the CP's were much smaller than 95%. We note that the differences in the estimates between the logit and C-log-log links might be due to the fact that the average hazards of less than 0.09 might not be smaller enough. To further examine this issue, we generated the discrete survival data with $\boldsymbol{\beta} = (0.2, 0.2)'$ and $\boldsymbol{\beta} = (0.02, 0.02)'$ under the ELTH model with the logit link, which yielded the average hazards less than 0.07 and 0.062 across 1,000 simulated datasets, respectively. These additional simulation results are given in Tables C1 and C2 of Appendix C in Electronic Supplementary Material. From these two tables, we now see that the estimates from the C-log-log link were almost the same as those from the true logit link.

These results suggest that (i) the estimates under the ELTH model with the C-log-log link are similar to those under the ELTH model with the logit link when the hazards are small although one is an asymmetric link while another link is symmetric; (ii) the ELTH models with probit and t_4 links fit the discrete survival data generated from the ELTH model with the logit link poorly; and (iii) fitting the discrete time survival data as the continuous time survival data can lead to a substantial bias in the estimation of regression coefficients.

The boxplots in Fig. 2 show the AIC differences between the ELTH models with the C-log-log, probit, and t_4 links and the true ELTH model with the logit link. Note that the BIC differences are exactly the same as the AIC differences in this case. The mean AIC differences were 0.10 and 0.17 between the C-log-log and logit links, 1.18 and 2.76 between the probit and logit links, and 1.64 and 3.02 between the t_4 and logit links for $n = 1000$ and $n = 2000$, respectively. Over 78 and 79% of the AIC differences between the probit and logit links and between the t_4 and logit links were above 0 for $n = 2000$. The AIC differences between the C-log-log and logit links were not noticeable, which empirically confirms our Empirical Observation. As discussed in Sect. 2.1, the discrete time hazard is the one under the ELTH with the C-log-log link if the data are generated from the continuous time Cox model. Therefore, ignoring the discreteness of observed survival data will lead to biased estimates of regression coefficients. We also note that if $F_1(u) = \exp(u)/\{1+\exp(u)\}$, then $\lim_{u \rightarrow -\infty} [F_1(c_1 u + c_2)/\Phi(u)] = \infty$ and $\lim_{u \rightarrow -\infty} [F_1(c_1 u + c_2)/F_{t_4}(u)] = 0$ for any $c_1 > 0$ and c_2 , where Φ and F_{t_4} denote the $N(0, 1)$ and t_4 cdf's. This explains the reason why the ELTH models with the probit and t_4 links did not fit well to the data generated under the logit link.

3.2 Results of simulation II

The means of AIC's and BIC's and frequencies of ranking each model as best for the four ELTH models corresponding to $J = 1, 5, 10$, and 15 under the logit link are reported in Table 3. The boxplots of the AIC and BIC differences between each of $J = 1, 10, 15$ and $J = 5$ (true) are shown in Fig. 3. From Table 3, we see that the true model with $J = 5$ had the smallest means of AIC, while the ELTH model with $J = 1$ had the smallest means of BIC; and AIC selected the true model 766 and 849 times out of 1000 simulations, corresponding to 76.6 and 84.9% powers, for $n = 1000$ and $n = 2000$, respectively, while BIC selected the true model only 88 and 406 times out of 1000 simulations for $n = 1000$ and $n = 2000$, respectively. These results are not surprising since BIC penalizes the dimension more severely than AIC. From Fig. 3, we see that all of the three boxes in Plot (a) or Plot (b) are above 0 while the whole box for $J = 1$ in Plot (c) and more than half of the box for $J = 1$ in Plot (d) are below 0. The findings from these boxplots are consistent with those in Table 3.

Similar to Simulation I, we also compare the ELTH models with four different links for the piecewise baseline hazard with $J = 5$. Fig. 4 shows the boxplots of AIC differences between the C-log-log, probit, and t_4 links and the true logit link. Again, the BIC differences are exactly the same as the AIC differences when we only compare the different links for a given value of J . The boxplots in this figure are very similar to those in 2.

4 Analysis of the SEER breast cancer data

For the subset of the SEER breast cancer data discussed in Sect. 1, among the 2096 female subjects, 1222 (58.3%) died after a median follow-up of 68 months (interquartile range (IQR): (27, 97.5) months) from the time of diagnosis. Based on Althuis et al. (2004) and Ries and Eisner (2007), the covariates considered in our analysis include age, tumor size (Size), extension of the primary tumor (Ext), positive lymph node involvement (PN), estrogen receptor status (ER), progesterone receptor status (PR), race, grade indicator, and treatment indicators. For the continuous covariates, the medians and IQRs of age, Size, and Ext were 55 and (45, 68) years, 60 and (40, 75) millimeters, and 15 and (11, 20) millimeters, respectively. The variables PN, ER, PR, race, and grade are binary covariates, taking value 0 or 1. Specifically, PN = 1 if the number of highest involved positive lymph nodes ≥ 5 ; ER = 1 if estrogen status is positive; PR = 1 if progesterone status is positive; race = 1 if not black; and grade = 1 if grade III. The treatment indicators include two binary variables, radiation (Rad) and Surgery (Surg), where Rad = 1 if radiation therapy received and Surg = 1 if surgery performed. The numbers of patients with the value 1 were 1071, 1118, 912, 1723, 1987, 1296, and 2025, respectively, for PN, ER, PR, race, grade, Rad, and Surg.

For this study cohort, we calculate a naïve estimate of the hazard at time t_k as

$$\hat{h}(k) = \frac{\text{number of events at time } t_k}{\text{number of subjects at risk at time } t_k}.$$

A plot of $\hat{h}(k)$ versus t_k is shown in Fig. 5. From this figure, we observe that all of the estimated hazards are less than 0.02.

We first obtain the maximum likelihood estimates (ESTs), the standard errors (SEs), and the p values of the parameters under the ELTH models with linear α_k given in (2.21) under the logit, C-log-log, probit, t_4 , and standard Laplace links. The results are given in Appendix C in Electronic Supplementary Material of the online supplementary materials. From Tables C. 3 and C.4, we observe that (i) the estimates and p values under the logit and C-log-log links are almost the same and those under the probit and t_4 links are different; (ii) the difference between the estimates for intercept under the C-log-log and standard Laplace link is approximately $\log(2)$; and (iii) the “time” variable is highly significant and the estimate of the corresponding coefficient ψ_1 is less than 0 under all the five links. The first two results are the direct consequences of our Empirical Observation and the third result implies that the data are from a population with a cure fraction of breast cancer according to Theorem 4.

To determine the choice of links and the piecewise baseline hazard model (2.22), we first fit the ELTH models with logit, C-log-log, probit, and t_4 links for various values of J to the SEER breast cancer data. Table 4 shows the values of AIC and BIC by fitting the ELTH models with piecewise baseline hazards under the above four links. We observe that (i) the smallest AIC value is achieved at $J = 35$ under each of the four links; (ii) the t_4 link yields the smallest values of AIC and BIC; (iii) the ELTH model with $J = 35$ has the smallest BIC value for logit, C-log-log, and t_4 links while the ELTH model with $J = 30$ has the smallest BIC value under the probit link; and (iv) the ELTH model with $J = 35$ fits the data much

better than the one with $J = 1$. From Table 4, we see that the values of AIC and BIC under the t_4 link are 4611.28 and 4865.43 for $J = 35$ and 5008.36 and 5076.13 for $J = 1$, respectively, and the results under other three links are similar.

Since the right tail portion of the hazard function is typically more flat than the portion at earlier time, it is expected that the right tail of the hazard should be more parsimonious. Thus, the second stage of our selection procedure for α_k is to further reduce the number of pieces from the right tail of the hazard in (2.22) under the “best” ELTH models with $J = 35$. Specifically, instead of (2.22), we consider

$$\alpha_k = \psi_0 + \sum_{j=1}^{J-2-J^*} \psi_j I\{s_j \leq t_k < s_{j+1}\} + \psi_{J-1-J^*} (t_k - s_{J-1-J^*})_+, \quad (4.1)$$

where $J^* \geq 0$ such that $J - 1 - J^* > 0$. We note that (4.1) with $J^* = 0$ reduces to (2.22) and a larger value of J^* yields a more parsimonious model. We choose the value of J^* according to the AIC or BIC criterion. Table 5 shows the AIC values for various choices of J^* . We see from Table 5 that the smallest AIC value is obtained at $J^* = 5$ under the logit, C-log-log, and t_4 links. Under the logit, C-log-log, probit, and t_4 links, the AIC values are 4610.85, 4610.57, 4618.26, and 4605.74, respectively, for $J^* = 5$, and 4614.21, 4613.96, 4620.26, and 4611.28, respectively, for $J^* = 0$. Similarly, the BIC value is also minimized at $J^* = 5$ for the logit, C-log-log, and probit links. In the remaining of the analysis of the SEER breast cancer data, we use α_k in (4.1) with $J = 35$ and $J^* = 5$.

The maximum likelihood estimates (ESTs), the standard errors (SEs), and the p values of the parameters under the ELTH models with $J = 35$ and $J^* = 5$ are reported in Table 6. The results under the standard Laplace link are given in Appendix C in Electronic Supplementary Material of the supplementary materials. The results for $J = 35$ and $J^* = 5$ are similar to those in Table C.4. The estimates of β and the p values under the logit are very close to those under the C-log-log link, which is expected according to our Empirical Observation. The time term, $(t - s_{29})_+$, is significant, with a p value ranging from 0.004 to 0.031, and the estimate of the corresponding coefficient is negative under all five links, indicating the cure fraction of the breast cancer data.

Under the ELTH model with the t_4 link, $J = 35$, and $J^* = 5$, we compute the estimated cure rates of $S(\infty|\mathbf{x}, \beta)$ and Fig. 6 shows the boxplots of these estimates stratified by the three treatment groups (surgery and radiation, only surgery, and other). Here, “only surgery” refers to “surgery without radiation” and a patient who had surgery might also receive treatments other than radiation. A summary of these cure rates is given in Table C.5 of Appendix C in Electronic Supplementary Material. From Fig. 6, we see that patients treated by both surgery and radiation clearly had higher cure rates than those treated by only surgery or other while patients treated by other had the lowest cure rates. For those patients who treated by both surgery and radiation, the minimum and maximum of the cure rates are 0.32 and 0.92 and the corresponding values of the covariate vector (Size, Ext, PN, ER, PR, grade, race, age) are (65, 50, 0, 0, 0, 1, 0, 84) and (59, 11, 1, 1, 1, 1, 1, 23), respectively. The information for other two treatment groups can be found in Table C.5 of Appendix C in Electronic Supplementary Material. The patient who had the worst cure rate of 0.012 in our

study cohort was 98 years old with grade IV breast cancer, treated by only surgery, and had a more than 6 of highest involved positive lymph nodes and negative estrogen and progesterone statuses. On the other hand side, the patient who had the highest cure rate of 0.92 was 23 years old with grade III breast cancer, and had a low number of highest involved positive lymph nodes and positive estrogen and progesterone statuses. These results are quite reasonable and clinically justifiable. In addition, Table C.6 of Appendix C in Electronic Supplementary Material shows the estimates of the parameters under the Cox regression model treating the discrete time survival data as the continuous survival time data. Compared Table C.6 to Table 6, we see some differences in the estimates of the parameters. Specifically, Size and race under the Cox model are more significant than those under the the ELTH models. Based on our simulation study, the estimates under the Cox model could be biased.

5 Discussion

As it is well known, the constant hazard model is the only model, which possesses the memoryless property for the continuous survival time. For the discrete survival time, under certain mild conditions, we can show that the ELTH model is memoryless if and only if a_k is constant, which implies that the baseline hazard function $h_0(k)$ is constant.

To compute the cure rate, we use the convergent criterion

$$k_r = \min \left\{ k: \left| \frac{S(k+1|\mathbf{x}, \boldsymbol{\beta}) - S(k|\mathbf{x}, \boldsymbol{\beta})}{S(k|\mathbf{x}, \boldsymbol{\beta})} \right| < 10^{-13} \right\}$$

and then approximate $S(\infty|\mathbf{x}, \boldsymbol{\beta})$ by $S(k_r|\mathbf{x}, \boldsymbol{\beta})$. The cure rate is useful in classifying breast cancer patients into different risk groups and aiding physicians to better treat patients.

In Sect. 2.8, we introduced the AIC and BIC criteria to determine the choice of links as well as the baseline hazards under the ELTH model. According to the simulation study in Sect. 3, AIC outperformed BIC. For the SEER breast cancer data, both AIC and BIC selected the same best model under logit, C-log-log, and t_4 links except for the probit link. Thus, AIC is a more desirable criterion in order to select a better fit model since BIC may over-penalize the dimension and sample size.

In Sect. 4, the knots, s_j 's, were specified by the Bi-Sectional Quantile Partition (BSQP) method proposed by Zhang et al. (2015), and the number (J) of intervals was determined by the AIC or BIC criterion. This approach can be extended by modeling s_j and J simultaneously via the dynamic models of Kim et al. (2007) and Wang et al. (2013). Such an extension improves flexibility of the model but also dramatically increases computation complexity. Also, in Sect. 4, we carried out a detailed analysis of the subset of the SEER breast cancer data. Our study cohort consisted of the female subjects who were at least 20 years old at diagnosis and were diagnosed with regional extension, grade III or IV, and stage III breast cancer between 1990 and 2003. As pointed out by an anonymous reviewer, there were missing observations among PN, ER, PR, race, Rad, and Surg. The missing percentages for PN, race, Rad, and Surg were small, ranging from 0.25 to 3.28%, while the

missing percentages for ER and PR were 9.98 and 10.41%. We simply excluded those subjects with missing values in any of these covariates and carried out a complete case analysis based on the 2096 subjects. However, our proposed methodology can be extended to allow the inclusion of subjects who had missing values in PN, ER, PR, race, Rad, and Surg. Following Ibrahim et al. (2012), such an extension requires the EM algorithm to carry out the data analysis, which can well be considered as a future research topic.

Following Cheng et al. (1995) and Banerjee et al. (2007), the proposed ELTH model may be extended as follows. Instead of (2.13), we take the transformation directly on the survival function as

$$F^{-1}(S(k|\mathbf{x}, \boldsymbol{\beta})) = \alpha_k + \mathbf{x}'\boldsymbol{\beta}. \quad (5.1)$$

A nice feature of this new formulation is that it directly connects the models discussed in Cheng et al. (1995) and Banerjee et al. (2007) for continuous survival data. However, (2.13) may be more attractive than (5.1) in the sense that (i) the parameters α_k and $\boldsymbol{\beta}$ in (2.13) have a better interpretation, especially under the logit link, since the hazard function under (5.1) becomes $h(k|\mathbf{x}, \boldsymbol{\beta}) = 1 - F(\alpha_k + \mathbf{x}'\boldsymbol{\beta})/F(\alpha_{k-1} + \mathbf{x}'\boldsymbol{\beta})$; (ii) (2.13) is more computationally convenient the existing computer software such as SAS procedures LOGISTIC and GENMOD can be directly used to fit discrete survival data; and (iii) the link F^{-1} in (5.1) may play a more substantial role than the one in (2.13) since $S(k|\mathbf{x}, \boldsymbol{\beta})$ can be large even when $h(k|\mathbf{x}, \boldsymbol{\beta})$ is small. Nevertheless, the potential usefulness of this extension needs to be further examined.

Throughout the paper, we only consider the logit, C-log-log, probit, t_4 , and standard Laplace links. A partial reason is that logit, C-log-log, and probit are three most popular links for modeling categorical data in practice. In addition, as described in our Empirical Observation, when the hazard is small, it is not a matter of choosing a skewed link or a symmetric link but the heavyness of the left tail of the link matters most for modeling discrete time survival data. Due to this reason, we consider the t_4 link as it has a heavier tail than the logit, C-log-log, and probit links. There are many other links available in the literature, including those proposed by Chen et al. (1999), Kim et al. (2007), and Jiang et al. (2013). However, the usefulness of these links in modeling discrete time survival data is not clear. This issue as well as the aforementioned extensions are currently under investigation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors wish to thank the Editor-in-Chief, the Associate Editor and two referees for helpful comments and suggestions, which have greatly improved the paper.

Funding Dr. Chen's research in this study was partially supported by NIH grants #GM 70335 and #P01 CA142538.

References

- Adeniji AK, Belle AH, Wahed AS. Incorporating diagnostic accuracy into the estimation of discrete survival function. *J Appl Stat.* 2014; 41:60–72.
- Althuis MD, Fergenbaum JH, Garcia-Closas M, Brinton LA, Madigan MP, Sherman ME. Etiology of hormone receptor-defined breast cancer: a systematic review of the literature. *Cancer Epidemiol Biomark Prev.* 2004; 13:1558–1568.
- Allison PD. Discrete-time methods for the analysis of event histories. *Sociol Methodol.* 1982; 13:61–98.
- Banerjee T, Chen MH, Dey DK, Kim S. Bayesian analysis of generalized odds-rate hazards models for survival data. *Lifetime Data Anal.* 2007; 13:241–260. [PubMed: 17401683]
- Biggeri L, Bini M, Grilli L. The transition from university to work: a multilevel approach to the analysis of the time to obtain the first job. *J R Stat Soc Ser A.* 2001; 64:293–305.
- Brown WJ, Steele F, Golalizadeh M, Green MJ. The use of simple reparameterizations to improve the efficiency of Markov Chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. *J R Stat Soc Ser A.* 2009; 172:579–598.
- Chen MH, Shao QM. Propriety of posterior distribution for dichotomous quantal response models. *Proc Am Math Soc.* 2001; 129:293–302.
- Chen MH, Dey DK, Shao QM. A new skewed link model for dichotomous quantal response data. *J Am Stat Assoc.* 1999; 448:1172–1186.
- Chen K, Jin Z, Ying Z. Semiparametric analysis of transformation models with censored data. *Biometrika.* 2002; 89:659–668.
- Cheng SC, Wei LJ, Ying Z. Analysis of transformation models with censored data. *Biometrika.* 1995; 82:835–845.
- Cox DR. Regression models and life-tables. *J R Stat Soc Ser B.* 1972; 34:187–202.
- Czado C, Santner TJ. The effect of link misspecification on binary regression inference. *J Stat Plan Inference.* 1992; 33:213–231.
- Efron B. Logistic regression, survival analysis, and the Kaplan-Meier curve. *J Am Stat Assoc.* 1988; 83:414–425.
- Fairchild AJ, Abara WE, Gottschall AC, Tein JY, Prinz RJ. Improving our ability to evaluate underlying mechanisms of behavioral onset and other event occurrence outcomes: A discrete-time survival mediation model. *Eval Health Prof.* 2013 doi:10.1177/0163278713512124.
- Grilli L. The random-effects proportional hazards model with grouped survival data: a comparison between the grouped continuous and continuation ratio versions. *J R Stat Soc Ser A.* 2005; 168:83–94.
- Ibrahim JG, Chu H, Chen M-H. Missing data in clinical studies: issues and methods. *J Clin Oncol.* 2012; 30(26):3297–3303. [PubMed: 22649133]
- Jiang X, Dey DK, Prunier R, Wilson AM, Holsinger KE. A new class of flexible link functions with application to species co-occurrence in cape floristic region. *Ann Appl Stat.* 2013; 7:2180–2204.
- Kalbfleisch JD, Prentice RL. Marginal likelihood based on Cox's regression and life model. *Biometrika.* 1973; 60:267–278.
- Kim S, Chen MH, Dey DK, Gamerman D. Bayesian dynamic models for survival data with a cure fraction. *Lifetime Data Anal.* 2007; 13:17–35. [PubMed: 17136621]
- Kim S, Chen MH, Dey DK. Flexible generalized t -link models for binary response data. *Biometrika.* 2008; 95:93–106.
- Manda S, Meyer R. Age at first marriage in Malawi: a Bayesian multilevel analysis using a discrete time-to-event model. *J R Stat Soc Ser A.* 2005; 68:439–455.
- Muthen B, Masyn K. Discrete-time survival mixture analysis. *J Educ Behav Stat.* 2005; 30:27–58.
- Nam CW, Kim TS, Park NJ, Lee HK. Bankruptcy prediction using a discrete-time duration model incorporating temporal and macroeconomic dependencies. *J Forecast.* 2008; 27:493–550.
- Nguyen VQ, Gillen DL. Robust inference in discrete hazard models for randomized clinical trials. *Lifetime Data Anal.* 2012; 8:446–469. [PubMed: 22810273]

- Obeysekara, PT. Unpublished Ph.D. Dissertation. University of Connecticut; 2013. Host selection of spring Tiphia (*Tiphia vernalis*) and summer Tiphia (*Tiphia popillivora*), natural enemies of Japanese and oriental beetles..
- Prentice RL, Gloeckler LA. Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*. 1978; 34:57–67. [PubMed: 630037]
- Prentice RL, Kalbfleisch JD. Mixed discrete and continuous Cox regression model. *Lifetime Data Anal.* 2003; 9:195–210. [PubMed: 12735496]
- Ries, L.; Eisner, M. SEER survival monograph: cancer survival among adults: U.S. SEER Program, 1988–2001. Lynn, A.; Gloeckler, R.; Milton, PE., editors. National Cancer Institute, SEER Program; Bethesda, MD: 2007. NIH Pub. No. 07–6215
- Steele F. A discrete-time multilevel mixture model for event history data with long-term survivors, with an application to an analysis of contraceptive sterilization in Bangladesh. *Lifetime Data Anal.* 2003; 9:155–174. [PubMed: 12735494]
- Singer JD, Willet JB. It's about time: using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*. 1993; 18:155–195.
- Stewart WH, Pierce DA. Efficiency of Cox's model in estimating regression parameters with grouped survival data. *Biometrika*. 1982; 69:539–545.
- Stukel TA. Generalized logistic models. *J Am Stat Assoc*. 1988; 83:426–431.
- Wang X, Chen MH, Yan J. Bayesian dynamic regression models for interval censored survival data with application to children dental health. *Lifetime Data Anal.* 2013; 19:297–316. [PubMed: 23389549]
- Wulczyn F, Chen L, Courtney M. Family reunification in a social structural context. *Child Youth Serv Rev.* 2011; 33:424–430.
- Zhang, D.; Chen, MH.; Ibrahim, JG.; Boye, ME.; Shen, W. *Journal of Statistical Software*. 2015; in press.. JMFIt: A SAS Macro for Joint Models of Longitudinal and Survival Data..
- Zhao X, Zhou X. Discrete-time survival models with long-term survivors. *Stat Med*. 2008; 27:1261–1281. [PubMed: 17676708]

Fig. 1.

The plots of the cumulative distribution functions corresponding to the probit, logit, C-log-log, t_4 , and standard Laplace links, where the whole cdf curve is shown in (a) and the enlarged portion of cdf over $(-6, -3)$ is shown in (b)

Fig. 2. Boxplots of AIC difference between each of the C-log-log, probit, and t_4 links and the logit link with the linear baseline hazard for α_k for $n = 1000$ and $n = 2000$

Fig. 3.

Boxplots of AIC difference between each of the C-log-log, probit, and t_4 links and the logit link with the piecewise baseline hazard for α_k with $J = 5$ for $n = 1000$ and $n = 2000$

Fig. 4.

Boxplots of AIC difference between each of the C-log-log, probit, and t_4 links and the logit link with the piecewise baseline hazard for α_k with $J = 5$ for $n = 1000$ and $n = 2000$

Fig. 5.
The plot of the estimated hazard versus time for the SEER breast cancer data

Fig. 6. Boxplots of the estimated cure rates stratified by treatment (surgery and radiation, only surgery, and others)

Table 1

Two small discrete survival data

Data	<i>i</i>	<i>t_{k_i}</i>	δ_i	<i>x_{i1}</i>	<i>x_{i2}</i>
<i>D</i> ₁ (<i>n</i> = 3)	1	1	1	0	0
	2	2	1	0	1
	3	1	1	1	1
<i>D</i> ₂ (<i>n</i> = 4)	1	1	1	0	0
	2	2	1	0	1
	3	1	1	1	1
	4	3	1	1	0

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2Estimates of β under the ELTH and Cox models

<i>n</i>	Parameter	Model	EST	ASE	SSD	RMSE	CP
1000	$\beta_1 = 0.5$	Logit (true)	0.499	0.041	0.040	0.040	0.951
		C-log-log	0.479	0.039	0.038	0.044	0.924
		Probit	0.238	0.020	0.019	0.263	0
		t_4	0.402	0.033	0.033	0.103	0.149
		Cox	0.459	0.039	0.036	0.055	0.820
	$\beta_2 = 0.5$	Logit (true)	0.498	0.077	0.079	0.079	0.947
		C-log-log	0.478	0.074	0.076	0.079	0.936
		Probit	0.236	0.036	0.038	0.266	0
		t_4	0.402	0.063	0.064	0.117	0.658
		Cox	0.459	0.074	0.073	0.083	0.920
2000	$\beta_1 = 0.5$	Logit (true)	0.501	0.029	0.029	0.029	0.949
		C-log-log	0.480	0.027	0.027	0.034	0.887
		Probit	0.238	0.014	0.014	0.262	0
		t_4	0.403	0.023	0.023	0.100	0.021
		Cox	0.460	0.027	0.026	0.047	0.696
	$\beta_2 = 0.5$	Logit (true)	0.500	0.055	0.055	0.055	0.955
		C-log-log	0.480	0.052	0.053	0.056	0.931
		Probit	0.238	0.026	0.026	0.264	0
		t_4	0.404	0.045	0.045	0.106	0.418
		Cox	0.462	0.052	0.051	0.064	0.889

Table 3

Means of AIC and BIC differences and frequencies of ranking each model as best based on AIC and BIC

	AIC				BIC			
	<i>n</i> = 1000		<i>n</i> = 2000		<i>n</i> = 1000		<i>n</i> = 2000	
	Mean	Frequency	Mean	Frequency	Mean	Frequency	Mean	Frequency
<i>J</i> = 1	6012.7	151	12,011.1	29	6032.3	912	12,033.5	594
<i>J</i> = 5	6006.7	766	11,995.7	849	6041.1	88	12,034.9	406
<i>J</i> = 10	6011.9	66	12,000.6	91	6070.0	0	12,067.3	0
<i>J</i> = 15	6016.5	17	12,004.9	31	6098.9	0	12,099.5	0

Table 4

AICs and BICs for the SEER breast cancer data

J	Link	Log-likelihood	AIC	BIC
1	Logit	-2492.48	5008.96	5076.74
	C-log-log	-2492.33	5008.66	5076.44
	Probit	-2497.65	5019.30	5087.07
	t_4	-2492.18	5008.36	5076.13
15	Logit	-2368.99	4787.99	4929.18
	C-log-log	-2368.83	4787.65	4928.85
	Probit	-2373.53	4797.07	4938.26
	t_4	-2363.71	4777.42	4918.62
30	Logit	-2281.89	4643.78	4869.69
	C-log-log	-2281.81	4643.62	4869.54
	Probit	-2283.95	4647.90	4873.81
	t_4	-2281.64	4643.29	4869.20
35	Logit	-2262.10	4614.21	4868.36
	C-log-log	-2261.98	4613.96	4868.11
	Probit	-2265.13	4620.26	4874.41
	t_4	-2260.64	4611.28	4865.43
40	Logit	-2260.65	4617.31	4888.40
	C-log-log	-2260.54	4617.09	4888.18
	Probit	-2263.55	4623.09	4894.18
	t_4	-2259.39	4614.77	4885.87

Table 5

AICs and BICs for reduction of the number of pieces for the piecewise baseline hazard under the ELTH model with $J = 35$

J^*	Link	Log likelihood	AIC	BIC
1	Logit	-2264.06	4616.12	4864.62
	C-log-log	-2263.95	4615.89	4864.39
	Probit	-2266.84	4621.68	4870.19
	t_4	-2262.98	4613.96	4862.46
3	Logit	-2263.57	4611.15	4848.36
	C-log-log	-2263.46	4610.91	4848.12
	Probit	-2266.52	4617.04	4854.25
	t_4	-2262.29	4608.58	4845.78
5	Logit	-2265.42	4610.85	4836.76
	C-log-log	-2265.29	4610.57	4836.48
	Probit	-2269.13	4618.26	4844.17
	t_4	-2262.87	4605.74	4831.65
6	Logit	-2269.64	4617.27	4837.54
	C-log-log	-2269.48	4616.96	4837.22
	Probit	-2273.95	4625.89	4846.16
	t_4	-2265.73	4609.46	4829.73
8	Logit	-2285.29	4642.59	4845.91
	C-log-log	-2285.14	4642.28	4845.60
	Probit	-2289.51	4651.01	4854.33
	t_4	-2282.11	4636.22	4839.54

Table 6Estimates of the parameters under the ELTH models for $J = 35$ and $J^* = 5$ for the SEER breast cancer data

Variable	EST	SE	P value	EST	SE	P value
Logit Link			C-log-log Link			
Intercept	-5.791	0.411	<0.001	-5.794	0.408	<0.001
Size	0.002	0.001	0.050	0.002	0.0010	0.049
Ext	0.010	0.003	<0.001	0.010	0.003	<0.001
PN	-0.505	0.106	<0.001	-0.502	0.106	<0.001
ER	-0.561	0.147	<0.001	-0.558	0.146	<0.001
PR	-0.172	0.151	0.255	-0.172	0.150	0.251
grade	-0.278	0.218	0.203	-0.277	0.217	0.202
race	-0.274	0.130	0.036	-0.273	0.129	0.035
Surg	-0.634	0.218	0.004	-0.630	0.215	0.003
Rad	-0.626	0.107	<0.001	-0.622	0.107	<0.001
age	0.024	0.004	<0.001	0.024	0.004	<0.001
$(t - s_{29})_+$	-0.019	0.008	0.015	-0.019	0.008	0.015
Probit Link			t_4 Link			
Intercept	-2.720	0.152	<0.001	-5.335	0.477	<0.001
Size	0.001	0.0004	0.067	0.002	0.001	0.039
Ext	0.004	0.001	<0.001	0.011	0.003	<0.001
PN	-0.180	0.038	<0.001	-0.578	0.129	<0.001
ER	-0.192	0.053	<0.00	-0.684	0.173	<0.001
PR	-0.049	0.053	0.357	-0.272	0.184	0.140
grade	-0.108	0.078	0.168	-0.286	0.255	0.262
race	-0.099	0.048	0.038	-0.309	0.146	0.035
Surg	-0.240	0.089	0.007	-0.679	0.206	<0.001
Rad	-0.226	0.038	<0.001	-0.707	0.132	<0.001
age	0.009	0.001	<0.001	0.029	0.004	<0.001
$(t - s_{29})_+$	-0.005	0.002	0.031	-0.042	0.014	0.004