Published in final edited form as:

Stat Med. 2008 November 20; 27(26): 5421–5439. doi:10.1002/sim.3368.

'Smooth' inference for survival functions with arbitrarily censored data

Kirsten Doehler¹ and Marie Davidian^{2,*†}

¹Department of Mathematics and Statistics, University of North Carolina at Greensboro, Greensboro, NC 27402-6170, U.S.A.

²Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203, U.S.A.

SUMMARY

We propose a procedure for estimating the survival function of a time-to-event random variable under arbitrary patterns of censoring. The method is predicated on the mild assumption that the distribution of the random variable, and hence the survival function, has a density that lies in a class of 'smooth' densities whose elements can be represented by an infinite Hermite series. Truncation of the series yields a 'parametric' expression that can well-approximate any plausible survival density, and hence survival function, provided the degree of truncation is suitably chosen. The representation admits a convenient expression for the likelihood for the 'parameters' in the approximation under arbitrary censoring/truncation that is straightforward to compute and maximize. A test statistic for comparing two survival functions, which is based on an integrated weighted difference of estimates of each under this representation, is proposed. Via simulation studies and application to a number of data sets, we demonstrate that the approach yields reliable inferences and can result in gains in efficiency over traditional nonparametric methods.

Keywords

bootstrap; information criteria; integrated weighted difference; interval censoring; seminonparametric density representation; truncation

1. INTRODUCTION

A routine challenge in medical, public health, and other bioscience applications is to estimate the survival function for a time-to-event (survival) endpoint based on a sample of incomplete observations. For example, interest may focus on time to death or tumor recurrence for cancer patients; time to progression to AIDS in HIV-infected individuals; or time to attainment of a composite endpoint such as myocardial infarction, urgent target vessel revascularization, or death in cardiovascular disease patients. In many clinical trials, event times are either observed exactly or are right censored, where right censoring may result from an individual's event time having not occurred by the end of the study or because of loss to follow-up. In other situations, the event may be known only to have occurred in some time interval, such as between periodic follow-up visits; for example, an individual may not satisfy the criteria for disease progression at one visit but meet them at the next. Here, event times are interval censored. Other forms of censoring or truncation of

event times are also possible; Pan and Chappell [1,2] discuss a study in which event times are both left truncated and interval censored (see Section 6.3).

Time-to-event data are often not well-described by parametric models [3, sec. 2.1.3]; accordingly, there has been a ‘biostatistical tradition’ [3] of favoring non- or semiparametric approaches to inference. For inference on the survival function on the basis of a sample of n independent, possibly censored or truncated event times, the methods of choice are nonparametric. The Kaplan-Meier (KM) estimator [4] for the survival function is ubiquitous for right censored data, enjoying widespread availability in popular software and standard asymptotic theory (with $n^{-1/2}$ convergence rate) that yields reliable finite-sample inferences. Being completely nonparametric, the KM estimator is discrete, taking jumps at the observed event times. Under interval censoring, the nonparametric maximum likelihood (NPML) estimator for the survival function has been studied by numerous authors, e.g., [5]; see Reference [6] for an introduction. However, the NPML estimator has nonstandard large sample properties [7], which make derivation of standard error estimators and confidence intervals problematic, and, as noted by Pan [8], it tends to have a smaller number of jumps and larger jump sizes than the KM estimator, providing poor visualization of the true survival function.

Despite the emphasis on nonparametric methods, it often may be reasonable to assume that the distribution corresponding to the true survival function possesses a continuous, ‘smooth’ density [8,9,10,11,12,13,14]. Settings where event times are sufficiently discrete so as to disqualify this assumption would likely be readily apparent to the data analyst. When this assumption is realistic, many authors have used smoothing techniques to represent the density. For example, Kooperberg and Stone [10] and Kooperberg *et al.* [11] represented the logarithm of the assumed density using polynomial splines with linear tail constraints. Komárek *et al.* [12] used penalized B-splines to smooth the baseline density in an accelerated failure time model (AFT), yielding what they call a mixture of ‘basis Gaussian densities,’ an expression for the assumed density in terms of a fixed, relatively large number of normal densities. These authors focused mostly on estimation of the density itself rather than the survival function [10,11] or on estimation of the finite-dimensional regression parameters in the AFT model [12]. Others [13,14] have derived theoretical results. If applied to estimation of the survival function, these approaches yield smooth, rather than discrete, estimators. Pan [8] studied performance of the NPML, kernel-smoothed NPML [15,16], and log-spline density-based [10] survival function estimators for interval censored data and found that inference predicated on ‘smoothness’ assumptions enjoys efficiency gains over NPML when these assumptions are valid.

These results suggest that estimation of a survival function under smoothness assumptions deserves further study. Recently, Zhang and Davidian [17] proposed a general approach to semiparametric censored data regression when the assumption of a ‘smooth’ baseline survival density is justified. The approach is based on the ‘SemiNonParametric’ (SNP) density representation of Gallant and Nychka [18], which suggests the form of a density estimator that has been used successfully as a foundation for nonlinear time series analysis [19] and as a model for random effects when deviations from normality are suspected [20,21,22]. The method relies on a procedure for using the SNP representation that is particularly well-suited to approximating densities of time-to-event random variables. Zhang and Davidian [17], as did Reference [12], focused on performance of estimators for regression parameters and recovery of the shape of the underlying baseline density. In this paper, we instead study use of the SNP representation for obtaining smooth, rather than discrete, estimates of and making inference on a survival function. In Section 2, we review the SNP representation and in Section 3 show how it may be used to provide a convenient, flexible approximation to the survival function that is straightforward to implement. We

propose an SNP-based test statistic for comparing two survival functions based on two independent samples in Section 4. Simulation studies in Section 5 demonstrate performance of the methods, which are illustrated by application to data from several medical and public health studies in Section 6.

2. THE SNP DENSITY REPRESENTATION

Gallant and Nychka [18] studied a class $\mathcal{H}$ of q -dimensional ‘smooth’ densities, the mathematical details of which are complex but, practically speaking, includes densities that have sufficient derivatives that they do not exhibit ‘unusual’ behavior such as oscillations or jumps, but may be skewed, multimodal, or fat- or thin-tailed. When $q = 1$, $\mathcal{H}$ is rich enough to include almost any density that is a plausible model for a (possibly transformed, continuous) time-to-event random variable but excludes those that are unlikely to be reasonable descriptions. Gallant and Nychka [18] showed that a density $h \in \mathcal{H}$ may be expressed as an infinite Hermite series, the ‘standardized’ form of which is $h(z) = P_\infty^2(z)\psi_q(z)$ plus a small lower bound governing tail behavior, where $P_\infty(z)$ is an infinite polynomial in z ($q \times 1$), and $\psi_q(z)$ is a q -variate density with a moment generating function. The density ψ_q is the standard q -variate normal density in almost all published accounts; however, any ‘standardized’ density with a moment generating function leads to a valid representation. We refer to ψ_q as the *base density* of the representation; the support of $h(z)$ is the same as that of the base density.

If the analyst is willing to assume that a density of interest lies in $\mathcal{H}$, this representation may be used as the basis for inference. In this case, it is standard to approximate h by ignoring the lower bound and truncating the infinite polynomial P_∞ , which yields the so-called SNP estimator for h [19]. We focus on $q = 1$ and drop the subscript q ; see Reference [21] for an account for general $q \geq 1$. The approximation to h based on truncation may be written as

$$h_K(z) = P_K^2(z)\psi(z), \quad (1)$$

where now $P_K(z) = a_0 + a_1z + a_2z^2 + \dots + a_Kz^K$ is the K th order polynomial in (scalar) z depending on a $(d \times 1)$ vector of coefficients $\mathbf{a} = (a_0, a_1, \dots, a_K)^T$, $d = K + 1$.

To ensure that $h_K(z)$ integrates to 1, it was shown in Reference [21] that $\int h_K(z) dz = 1$ may be imposed ‘automatically’ by noting that this is the same as requiring that $E\{P_K^2(Z)\} = 1$, where Z has density ψ . Writing $P_K(Z) = \mathbf{a}^T \mathbf{U}$ for $\mathbf{U} = (1, Z, \dots, Z^K)^T$, this may be written as $\mathbf{a}^T E(\mathbf{U}\mathbf{U}^T) \mathbf{a} = \mathbf{a}^T \mathbf{A} \mathbf{a} = 1$. The positive definite $(d \times d)$ matrix $\mathbf{A}$ is easily calculated from the moments of the known, standardized density ψ . Because for $\mathbf{A}$ positive definite there exists $\mathbf{B}$ such that $\mathbf{A} = \mathbf{B}^2$, this reduces to requiring that $\mathbf{c}^T \mathbf{c} = 1$, where $\mathbf{c} = \mathbf{B}\mathbf{a}$. Thus, $\mathbf{c}$ lies on the half-unit sphere in $\mathbb{R}^d$, suggesting the spherical transformation $c_1 = \sin(\phi_1)$, $c_2 = \cos(\phi_1) \sin(\phi_2)$, $\dots$, $c_{d-1} = \cos(\phi_1) \cos(\phi_2) \dots \cos(\phi_{d-2}) \sin(\phi_{d-1})$, $c_d = \cos(\phi_1) \cos(\phi_2) \dots \cos(\phi_{d-2}) \cos(\phi_{d-1})$ for $-\pi/2 < \phi_\ell \leq \pi/2$, $\ell = 1, \dots, d-1$ [21]. It follows that (1) may be expressed in terms of $\boldsymbol{\phi} = (\phi_1, \dots, \phi_{d-1})^T$, which we emphasize henceforth by writing $h_K(z; \boldsymbol{\phi}) = P_K^2(z; \boldsymbol{\phi})\psi(z)$. See Web Appendix B of Reference [17] for examples of $\mathbf{A}$, $\mathbf{B}$, and representation of $\mathbf{a}$, and hence P_K , in terms of $\boldsymbol{\phi}$ when $\psi(z) = \varphi(z) = (2\pi)^{-1/2}e^{-z^2/2}$, the standard normal density, and $\psi(z) = \varepsilon(z) = e^{-z}$, the standard exponential density.

The order K of the polynomial in (1) controls the degree of flexibility for approximating the true density. When $K = 0$, $P_K^2(z) \equiv 1$, so there is no $\boldsymbol{\phi}$ and $h_K(\cdot) = \psi(\cdot)$, and the base density itself is a special case. For $K > 0$, K is often misinterpreted as the number of components in a mixture; rather, it dictates the extent of departure from the base density. The representation

$h_K(z, \phi)$ can capture a wide range of behavior and shapes with K as small as 1 or 2 [20,21,22,23], with $K \leq 4$ often sufficient to provide an excellent approximation.

3. ESTIMATION OF THE SURVIVAL FUNCTION

3.1. Representation of survival density and survival function

Let T be a time-to-event random variable having distribution with support on $(0, \infty)$, assumed to have density $f(t)$ and survival function $S(t) = P(T > t)$, $t > 0$. Our objective is inference on $S(t)$. Assume f lies in a class of ‘smooth’ densities whose elements may be approximated via SNP. We follow Zhang and Davidian [17] and propose an approach to estimation of $S(t)$ in which K and the base density are chosen adaptively.

Following the convention in other applications of taking the SNP base density to be standard normal, it is natural to assume that T satisfies

$$\log(T) = \mu + \sigma Z, \quad \sigma > 0, \quad (2)$$

where Z has density $h \in \mathcal{H}$ with support on $(-\infty, \infty)$ that can be approximated by $h_K(z)$ as in (1) with standard normal base density $\psi(\cdot) = \phi(\cdot)$. If $K = 0$, (2) reduces to assuming that T is lognormally distributed; for general fixed K , (2) leads to the approximation to $f(t)$

$$f_K(t; \theta) = \frac{1}{t\sigma} P_K^2 \left(\frac{\log t - \mu}{\sigma}; \phi \right) \varphi \left(\frac{\log t - \mu}{\sigma} \right), \quad t > 0, \quad \theta = (\mu, \sigma, \phi^T)^T. \quad (3)$$

Although (3) using the normal base density can capture a broad range of behavior for small K (≤ 4 , say), satisfactory approximation of some survival densities, e.g., the exponential and very skewed densities, requires larger K . Because computational burden increases with K , a more parsimonious approach is to assume that densities in the class in which f lies also admit another representation better suited to these conditions. In particular, assume that the density of Z in (2) is such that the density of $Z^* = e^Z$ may be approximated by (1) with base density $\psi(\cdot) = \varepsilon(\cdot)$, the standard exponential density; when $K = 0$, Z thus has an extreme value distribution [24, p. 32]. We then have the alternative approximation

$$f_K(t; \theta) = \frac{t^{(1/\sigma-1)}}{\sigma e^{\mu/\sigma}} P_K^2 \left\{ \left(\frac{t}{e^\mu} \right)^{1/\sigma}; \phi \right\} \varepsilon \left\{ \left(\frac{t}{e^\mu} \right)^{1/\sigma} \right\}, \quad t > 0, \quad \theta = (\mu, \sigma, \phi^T)^T. \quad (4)$$

When $K = 0$ in (4), this amounts to assuming that T has a Weibull distribution; when $K = 0$ and $\sigma = 1$, T has an exponential distribution.

For a specific representation of $f(t)$ defined by a fixed K and base density as in (3) or (4), the corresponding representation of $S(t)$ is straightforward. Under (2) with the normal base density, (3) leads to the approximation

$$S_K(t; \theta) = \int_{(\log t - \mu)/\sigma}^{\infty} P_K^2(z; \phi) \varphi(z) dz, \quad \theta = (\mu, \sigma, \phi^T)^T, \quad (5)$$

which is easily seen to be a linear combination of integrals of the form $I(k, c) = \int_c^{\infty} z^k \varphi(z) dz$ that obey the recursive relationship $I(k, c) = c^{k-1} \varphi(c) + (k-1)I(k-2, c)$, $k \geq 2$, where $I(0, c) = 1 - \Phi(c)$, $I(1, c) = \varphi(c)$, and $\Phi(\cdot)$ is the standard normal cumulative distribution function (cdf). Similarly, with the exponential base density, (4) yields the approximation

$$S_K(t; \theta) = \int_{(t/e^\mu)^{1/\sigma}}^{\infty} P_K^2(z; \phi) \varepsilon(z) dz, \quad \theta = (\mu, \sigma, \phi^T)^T, \quad (6)$$

where the integrals in (6) may be calculated using the recursion

$$I(k, c) = \int_c^{\infty} z^k e^{-z} dz = c^k \varepsilon(c) + kI(k-1, c), \quad k \geq 1, \quad \text{and } I(0, c) = e^{-c}.$$

Thus, under the assumption that T can be represented as (2), $f(t)$ and $S(t)$ may be approximated straightforwardly in terms of a finite-dimensional ‘parameter’ θ for fixed K and base density (standard normal or exponential), where $p = \dim(\theta) = K + 2$, which is modest when $K \leq 4$. When $K = 0$, θ reduces to $(\mu, \sigma)^T$.

3.2. Maximizing the likelihood function

Given a fixed K -base density approximation to $f(t)$ and $S(t)$ depending on θ as in (3) or (4), the loglikelihood for θ , and hence for $f(t)$ and $S(t)$, based on a sample of observed data under standard assumptions may be formulated immediately. For the case of independent right censoring, with C a positive random variable representing potential (right) censoring time, let $V = \min(T, C)$ and $\Delta = I(T \leq C)$ be the observed data. Under the usual assumption that T and C are independent, given a sample of independent and identically distributed (iid) pairs (V_i, Δ_i) , $i = 1, \dots, n$, the loglikelihood is

$$\ell_K(\theta) = \sum_{i=1}^n [\Delta_i \log\{f_K(V_i; \theta)\} + (1 - \Delta_i) \log\{S_K(V_i; \theta)\}]. \quad (7)$$

Similarly, if T is subject to independent right or interval censoring, we observe iid (L_i, R_i, Δ_i) , $i = 1, \dots, n$, say, where T_i is either known only to lie in the interval $[L_i, R_i]$ or is right censored at L_i (set $R_i = \infty$), in which case $\Delta_i = 0$; else, T_i is observed exactly, in which case $L_i = R_i = T_i$ and $\Delta_i = 1$. In this case, the corresponding loglikelihood is

$$\ell_K(\theta) = \sum_{i=1}^n [\Delta_i \log\{f_K(L_i; \theta)\} + (1 - \Delta_i) \log\{S_K(L_i; \theta) - S_K(R_i; \theta)\}]. \quad (8)$$

It should be clear that the loglikelihood for θ under any arbitrary censoring/truncation pattern, e.g., left truncation and interval censoring [1], is straightforward. In particular, $S_K(t; \theta)$ in (5) or (6) is trivially calculated recursively in a closed form, the former requiring only a standard numerical routine to evaluate the normal cdf.

Under (7), (8), or other suitable loglikelihood, for fixed K and base density, $S(t)$ may be estimated for any t by $S_K(t; \hat{\theta})$, where $\hat{\theta}$ maximizes $\ell_K(\theta)$. Because θ has fixed dimension for given K and base density, this may be accomplished using standard routines for finite-dimensional optimization; we have used successfully the SAS IML optimizer nlqp [25].

As with any optimization problem, care must be taken to ensure that the global maximum is attained. This can be achieved by carrying out the optimization over a ‘grid’ of starting values and selecting the ‘solution’ $\hat{\theta}$ yielding the largest value of $\ell_K(\hat{\theta})$, as in other applications of SNP [17, 19, 20]. To define a suitable grid of θ values, we have used the following approach, described first for right censored data. Assuming (2), $E\{\log(T)\} = \mu + \sigma E(Z)$ and $\text{var}\{\log(T)\} = \sigma^2 \text{var}(Z)$; thus, letting $m_{\log(T)}$ and $s_{\log(T)}^2$ be the sample mean and variance of the uncensored $\log(V_i)$ (those with $\Delta_i = 1$), respectively, we have

$$m_{\log(T)} \approx \mu + \sigma E(Z), \quad s_{\log(T)}^2 \approx \sigma^2 \text{var}(Z). \quad (9)$$

For given K and base density, any fixed value of ϕ fully determines $E(Z)$ and $\text{var}(Z)$. Thus, we suggest defining a grid over ϕ by letting each component ϕ_ℓ , $\ell = 1, \dots, d-1$, take on each of a set of predetermined values in the interval $(-\pi/2, \pi/2)$. For each possible combination of these $d-1$ values, the resulting ϕ determines $E(Z)$ and $\text{var}(Z)$, whereupon corresponding starting values for $(\mu, \sigma)^T$ are obtained by ‘solving’ (9), leading to a starting value for θ . When $K = 0$, there is no ϕ , and this is trivial; when $K = 1$, we recommend taking ϕ in $[-1.5, 1.5]$, as $\pi/2 \approx 1.57$, in increments of 0.2. For $K = 2$, $\phi = (\phi_1, \phi_2)^T$, and we take $\phi_\ell = -1.5, -0.5, 0.5, 1.5$; as K and hence $\dim(\phi)$ increases, we take a coarser grid in each component to maintain a feasible number of optimizations to be performed. For interval censored data, we use an analogous procedure, where interval censored event times are replaced by the midpoints of their respective intervals in the calculation of $m_{\log(T)}$ and $s_{\log(T)}^2$.

3.3. Choosing the degree of truncation and base density

The foregoing developments apply to a fixed K and base density representation. Although a fixed representation may provide sufficient flexibility to approximate many densities, an objective, adaptive approach is preferable. Following previous applications of SNP [17,19,20,21,22], we select the K -base density combination that provides the ‘best’ approximation to the true $f(t)$ and $S(t)$ by inspection of the values of a suitable information criterion over the possible combinations of K and base density, where $K = 0, 1, \dots, K_{\max}$ for some maximum value $K_{\max}$. Taking $K_{\max} = 3$ or 4 is sufficient as long as both the standard normal and standard exponential base densities representations (3) and (4) are considered.

In situations not involving censoring and/or truncation, three criteria having the form of a penalized negative loglikelihood $-\ell_K(\hat{\theta}) + pC(n)$, say, $p = \dim(\theta)$, are routinely used with SNP [21], where small values are preferred. The Akaike Information Criterion (AIC) takes $C(n) = 1$, the Bayesian Information Criterion (BIC) takes $C(n) = \log(n)/2$, and the intermediate Hannan-Quinn (HQ) criterion takes $C(n) = \log\{\log(n)\}$. AIC tends to select models with larger p and BIC is conservative, while HQ is intermediate. With censored or otherwise incomplete data, Kooperberg and Clarkson [26, sec. 3] cautioned that dependence of $C(n)$ on n as in HQ and BIC misrepresents the disparate uncertainty in the observations. For right censored data, References [11] and [27] replaced n by n_d , the number of observed failures; for interval censored data, an analogous approach could be taken, where now n_d is the number of observations that are observed failures or known to be left or interval censored. We have investigated the relative performance of using n and n_d with the proposed SNP approach and have found that, unless n is very small or right censoring is very heavy, so n and n_d differ substantially, the same K -base density combination tends to be chosen by HQ and BIC using n or n_d at least 98% of the time. As found by other authors in complete data settings, HQ tends to strike the best balance between parsimony and flexibility. Based on these considerations, we use HQ with n_d replacing n in the empirical studies and applications in Section 5 and Section 6.

Repeated fitting for each K -base density combination, where each fit requires optimization over a grid of starting values to ensure the global maximum for that combination is reached, may seem computationally prohibitive. Our implementation in SAS proc iml using the nlpq optimizer [25] requires only a few seconds to complete this entire process and identify the ‘best’ model for practical sample sizes of $n = 200$ or greater.

Note that if $K = 0$ is selected, this may be interpreted as supporting the parametric model implied by the corresponding base density. Thus, this procedure also provides an informal means of evaluating whether the data contain evidence for departures from the (parametric) lognormal or Weibull models.

3.4. Standard errors

Once the ‘best’ K -base density approximation is chosen, the estimator for $S(t)$ is the corresponding $S_K(t; \hat{\theta})$. To obtain pointwise standard errors and confidence intervals for $S(t)$ at any t , we consider two approaches. In the first, we follow other authors [17,20,21,22] and approximate the sampling variance of $S_K(t; \hat{\theta})$ by treating the chosen K and base density as fixed and known, even though they are chosen adaptively. That is, regarding $S_K(\cdot)$ for the chosen model as the loglikelihood of a prespecified parametric form, obtain an estimate of the inverse ‘information matrix’ (e.g., from the Hessian calculated by the optimizer) and use the delta method to derive an estimator for $\text{var}\{S_K(t; \hat{\theta})\}$ in the usual way. In other contexts, this has been shown empirically to work well in moderate-to-large samples. Alternatively, a nonparametric bootstrap in which (V_i, Δ_i) pairs are sampled with replacement B times may be used to obtain a standard error estimate. Assuming that $S_K(t; \hat{\theta})$ has approximately a normal sampling distribution with variance calculated by either method, pointwise $100(1-\alpha)\%$ Wald confidence intervals may be constructed as the estimate $\pm$ the normal critical value $\times$ the estimated standard error. To date, neither approach has formal theoretical justification; however, see Web Appendix A of Reference [17] for a review of the rationale for the delta method procedure. Use of the bootstrap has been proposed with other techniques (e.g., [8]) where theory is similarly lacking.

We evaluate the performance of both approaches empirically in Section 5.

4. TWO-SAMPLE TEST STATISTIC

A common objective is to compare survival functions for two populations, $S_1(t)$ and $S_2(t)$, say, on the basis of two independent samples of possibly censored event times of sizes n_1 and n_2 . That is, we wish to test $H_0 : S_1(t) = S_2(t), t \geq 0$, versus $H_1 : S_1(t) \neq S_2(t)$, where, for example, $S_1(t)$ and $S_2(t)$ characterize survival probabilities when a population is exposed to two different treatments. Via the approach of the last section, we may define independent estimators, in obvious notation, $S_{1,K_1}(t; \hat{\theta}_1)$ and $S_{2,K_2}(t; \hat{\theta}_2)$, where K and the base density are chosen separately for each group, and use these as the basis for a test statistic. Following References [8,28,29], we propose a test based on the integrated weighted difference

$$IWD = \int_0^{t_0} w(t) \{S_{1,K_1}(t; \hat{\theta}_1) - S_{2,K_2}(t; \hat{\theta}_2)\} dt, \quad (10)$$

where $w(t)$ is a weight function [28], and t_0 is a time point such as the length of the study or a time at which there is no longer information on survival times [28]. The simplest version of (10) takes $w(t) \equiv 1$. Under right censoring, to achieve stability, Reference [28, sec. 4] suggests taking instead $w(t) = \{C_1(t)C_2(t)\} / \{p_1C_1(t) + p_2C_2(t)\}$, where $C_j(t), j = 1, 2$, are the censoring survival functions in each group; and $p_j = \lim_{n \rightarrow \infty} (n_j/n), j = 1, 2$. Here, one may estimate $C_j(t), j = 1, 2$, by assuming that the censoring distributions have ‘smooth’ densities;

representing each using SNP to obtain estimates $C_{j,K_j}(t; \hat{\theta}_j), j = 1, 2$, say; and substituting these in $w(t)$ along with n_j/n for $p_j, j = 1, 2$. Alternatively, $w(t)$ may be chosen on subject-matter grounds [28]. Whether $w(t)$ is fixed or estimated, an estimator for the variance of IWD may be constructed using the delta method, assuming as in Section 3.4 that the relevant K -base combinations for the estimated survival (and censoring survival, where applicable) functions are fixed; instead, the bootstrap may be used. The test statistic is then (10) divided

by the square root of the estimated variance. In Section 5.3, we demonstrate that comparison of such statistics to standard normal critical values yields tests with reliable operating characteristics.

5. SIMULATION STUDIES

We have carried out numerous simulations under right and interval censored data scenarios with different sample sizes and true survival distributions to evaluate the performance of the SNP approach and to compare it to competing methods. We report on a subset of these here.

5.1. Inference on the survival function

We first summarize simulations involving independent uniform right censoring; results for exponential censoring were similar. We report results for four true distributions for T : Lognorm(0,1), the lognormal obtained by exponentiating draws from $\mathcal{N}(m, s^2)$, the normal distribution with mean m and variance s^2 , with $m = 0$, $s^2 = 1$; Weib(1.5,2.5), Weibull with shape and scale parameters 1.5 and 2.5; a distribution generated by exponentiating draws from a mixture of normal distributions, given in shorthand as $0.7\mathcal{N}(-1.1, 0.75^2) + 0.3\mathcal{N}(1.5, 0.75^2)$; and a mixture of Weibull and lognormal distributions given by $0.7 \text{ Weib}(2.5, 3.5) + 0.3 \text{ Lognorm}(-0.8, 0.8^2)$. The first two distributions correspond to familiar forms, so test the ability of SNP to capture ‘parametric’ behavior. The last two are more ‘unusual’ (asymmetric bimodal) and address performance of smoothing methods in situations where flexibility is required, and are representative of several such scenarios we have investigated. The censoring variable C was generated independently from a uniform distribution chosen to yield 30% or 50% right censoring in each case. We give results for $n = 400$ and comment on smaller n below. For each combination of distribution and censoring rate, 1000 Monte Carlo replications were performed.

For each data set, we obtained the SNP estimate of $S(t)$ by taking $K = 0, 1, 2, 3$ for each of the base density representations (1) and (4) and choosing the K -base density representation minimizing the HQ criterion as discussed in Section 3.4. Table I shows the number of times each combination was chosen for the four distributions. Not unexpectedly, for both censoring rates, the $K = 0$ /normal approximation is preferred for 80–90% of the data sets when $S(t)$ is lognormal, while the $K = 0$ /exponential is similarly preferred when $S(t)$ is Weibull. The ‘best’ representations for the mixture scenarios also gravitate to a specific K -base density combination for 30% censoring, with the normal (exponential) base density with $K = 2$ ($K = 3$) preferred for the exponentiated normal (lognormal-Weibull) mixture. For moderate to large n and moderate censoring, this tendency for a particular K -base density combination to emerge as the favored representation is commonplace, as has been seen in other uses of SNP. With 50% censoring, a preferred representation is less clear. Overall, allowing the choice of the two base densities affords considerable flexibility for approximating $S(t)$ with small K .

We also estimated $S(t)$ using the KM estimator; the estimator for $S(t)$ in Reference [10], implemented in the R function `oldlogspline` in the R packages `logspline`[30] and `polspline`[31]; and the estimator for $S(t)$ following from Reference [12], implemented in the R package `smoothSurv`[32]. The latter two estimators also involve smoothing, so are competing methods that would be expected to perform similarly to the SNP approach. For each estimator, we computed at several times t Monte Carlo relative bias; mean square error (MSE); and relative efficiency of estimator A to estimator B calculated as MSE for B divided by that for A . For SNP, we constructed 95% Wald confidence intervals for each $S(t)$, with standard errors calculated using the delta method as in Section 3.4 (SNP) and Greenwood’s formula [24, sec. 1.4.1] (KM); such analytic standard errors are not available from `oldlogspline` and `smoothSurv`. Table II shows results for estimation of $S(t)$ at the times t

where $S(t) = 0.8, 0.6, 0.4, 0.2$ for each true distribution. Relative biases were less than 3% for all estimators in all cases and are not shown.

We experienced considerable convergence issues with **oldlogspline** and had to experiment with strategies to obtain a high proportion of convergent data sets that could be automated in a simulation. After extensive testing, we settled on an approach using the default stepwise knot deletion option. If convergence was not achieved by **oldlogspline** automatically picking the optimal number of knots, we then tried a varying number of decreasing knots until convergence was successful. We resorted to using methods that did not involve stepwise knot deletion if convergence problems still arose. For some data sets, particularly for the exponentiated normal mixture scenario, convergence could not be achieved at all.

Confidence intervals based on the KM and SNP estimators under 30% censoring attain the nominal level in most cases and tend to exhibit slight under coverage at the same time points. Not unexpectedly, performance is degraded with 50% censoring at times near the upper censoring limit (beyond which no observed event times are available). For SNP, this is due to underestimation of the true sampling variation by the delta method when the available information on survival is scant. Under coverage becomes more serious with smaller sample sizes (not reported); for large samples, virtually all coverages are 0.95 for both 30% and 50% censoring, even when extrapolating modestly beyond observed event times for SNP. This suggests that SNP-based confidence intervals using the delta method should only be considered for situations with larger sample sizes where censoring is not heavy. In all cases, pointwise estimation of $S(t)$ via the SNP approach enjoys impressive efficiency gains relative to the KM estimator. Efficiency of the SNP relative to the **smoothSurv** estimator is either the same or shows modest gains when the true distribution involves the normal either directly or through a mixture, which is not surprising given the latter estimator's construction using basis Gaussian densities. Also not surprising are the larger gains in efficiency for SNP relative to the two smoothing estimators in the Weibull case, as SNP can represent this distribution exactly. Efficiency of the SNP relative to the **oldlogspline** estimator shows modest to large gains relative to the other distributions as well.

Figure 1 shows the simulation results graphically for each distribution under 30% censoring; plots for 50% censoring are similar. Unbiasedness of all estimators is evident. In the range of observed event times, the estimates are very similar, with the SNP, **oldlogspline**, and **smoothSurv** estimates looking like 'smooth' versions of the KM estimates. Although not readily apparent from the plots, the SNP estimates are less variable, reflected in the relative efficiency results in Table II.

For 'low information' situations with small n or heavy censoring, we investigated in additional simulations performance of confidence intervals based on standard errors obtained using a nonparametric bootstrap with $B = 100$ for the SNP, **oldlogspline**, and **smoothSurv** estimators; standard errors for the KM estimator were again computed via Greenwood's formula. Table III shows results for the same four true distributions, where each scenario involved 100 Monte Carlo data sets with $n = 100$ and 30% uniform censoring. SNP standard errors were also obtained using the delta method for comparison. Wald confidence intervals were constructed for all estimators using the indicated standard errors. Relative biases of all estimators are less than 3% in all cases, and the relative efficiencies of the SNP estimator to the KM, **oldlogspline**, and **smoothSurv** estimators are similar to those shown in Table II (not reported). Both the KM and SNP confidence intervals based on analytical (Greenwood or delta method) standard errors show a tendency for undercoverage in this small sample size, while the SNP intervals constructed using bootstrap standard errors achieve nominal coverage in most cases. Although these simulations are imprecise

owing to the small Monte Carlo sample size, they and others we have conducted suggest that use of the bootstrap in this way in ‘low information’ settings yields good results. Alternative bootstrap intervals (e.g., based on bootstrap percentiles) could also be used at the expense of much larger B .

We also carried out simulations under various patterns of right and interval censoring. We report here on scenarios involving the same four true distributions as above with approximately 20% right and 80% interval censored data generated according to the following general scheme. For each distribution, we chose five equally spaced, fixed ‘visit times’ $0 < s_1 < \dots < s_5$ to achieve this censoring pattern. For each subject, actual, random visit times were generated as draws from non-overlapping uniform distributions around each $s_j, j = 1, \dots, 5$. An event time T was then independently generated, and the interval $[L, R]$ for each subject was defined by the actual visit times between which T fell, with T right censored if it exceeded the final visit time. For each data set, we computed the SNP, **oldlogspline**, **smoothSurv**, and NPML estimates of $S(t)$, the latter calculated using the algorithm of Turnbull [5]. For each distribution, we took $n = 500$ and carried out 1000 Monte Carlo replications.

The numbers of data sets for which specific K -base density SNP representations were chosen were similar to those in Table I. Table IV and Figure 2 present results for estimation of $S(t)$. Relative biases were less than 2% for all estimators (not shown). Because no simple approach to calculate standard errors for the NPML estimator is available, nor are analytical estimates available from **oldlogspline** or **smoothSurv**, we present Monte Carlo coverage only for SNP-based Wald intervals using delta method standard errors along with relative efficiency defined as the MSE for the competing estimators divided by that for SNP.

SNP-based confidence intervals attain the nominal level with a tendency for undercoverage in some cases; small simulations similar to those in Table III show that bootstrap-based SNPWald confidence intervals can achieve acceptable performance in such cases, so may be preferred except in very large samples.

5.2. Comparison to parametric models

Given the ‘parametric’ form of the SNP estimator, it is natural to consider the extent to which it achieves performance approaching that of fitting the correct parametric model when the true distribution is a familiar one such as the lognormal or Weibull. Table II presents Monte Carlo efficiency of the SNP estimator relative to the estimator obtained by fitting the correct model using SAS **proc lifereg** [25], defined as the Monte Carlo MSE for the correct model fit divided by that for SNP, when the true distribution is lognormal or Weibull under 30% right censoring. Wald confidence interval coverages based on the correct model fits are similar to those for SNP and KM. Efficiency loss incurred by using SNP is modest at most time points; thus, an analyst contemplating such a parametric model can gain protection in the event that this assumption is incorrect while maintaining the potential for near-parametric performance.

5.3. Hypothesis testing

We carried out simulations of the operating characteristics of the SNP test statistic based on (10) under two scenarios, each with 1000 data sets and $n_1 = n_2 = 200$. We computed the SNP test statistic using delta method standard errors with $w(t) \equiv 1$ and $w(t)$ estimated as discussed following (10) as well as the standard logrank statistic. In all cases we took t_0 to be the minimum of the largest observed event times in each group. Each test was carried out at level 0.05. The SNP test statistics performed similarly, so we report results only for $w(t) \equiv 1$.

In the first scenario, with 30% uniform censoring in each group, under the null hypothesis, $S_1(t)$ and $S_2(t)$ corresponded to the Weibull with shape and scale parameters 1.5 and 2.5. The SNP and logrank tests both showed Monte Carlo rejection rates very close to 0.05. Under the alternative that $S_2(t)$ is Weibull with shape and scale 1.8 and 2.1, the SNP and logrank tests rejection rates were 0.65 and 0.72, respectively, while under the alternative that $S_2(t)$ is lognormal with location 0.1 and scale 1.4, the SNP and logrank rejection rates were 0.66 and 0.61. In the second scenario, with 20% and 40% uniform censoring in groups 1 and 2, under the null, $S_1(t)$ and $S_2(t)$ were both standard lognormal, and the SNP and logrank rejection rates were again 0.05. Under the alternative that $S_2(t)$ is lognormal with location and scale 0.2 and 1.2, the SNP and logrank rejection rates were 0.50 and 0.47. Overall, all of our simulations suggest that the SNP-based test achieves the nominal level and is competitive with the logrank test under most conditions, routinely exceeding the power of the logrank test when the alternative involves $S_1(t)$ and $S_2(t)$ from different distributional families. Greater power is likely possible by choosing $w(t)$ to emphasize anticipated alternatives [28, sec. 4].

6. APPLICATIONS

We further demonstrate the proposed methods by application to several familiar data sets. In each case, the SNP test statistic based on (10) was computed using weights $w(t) \equiv 1$.

6.1. ACTG 175

AIDS Clinical Trials Group Protocol 175 (ACTG 175) [33] randomized HIV-infected subjects to four different antiretroviral regimens in equal proportions: zidovudine (ZDV) monotherapy, ZDV + didanosine (ddI), ZDV+zalcitabine, and ddI monotherapy. We follow Reference [34] and consider two groups: ZDV monotherapy, with $n_1 = 617$ subjects, and the other three groups combined, with $n_2 = 1847$ subjects. The time-to-event endpoint was time to progression to AIDS or death, which was right censored for 68% and 80% of the participants in the ZDV alone and combined treatment groups, respectively.

The SNP procedure was applied to each treatment group. For the ZDV monotherapy group, HQ selected the standard normal base density with $K_1 = 0$ and, for the combined group, the standard exponential base density with $K_2 = 2$. SNP survival function estimates for each group are shown in Figure 3 with the KM estimates superimposed. The square of the SNP test statistic based on (10) and the delta method is 39.9 compared to the logrank statistic of 47.2; using the bootstrap yields a squared statistic of 62.8. Comparison of all to the χ^2_1 distribution indicates very strong evidence of a difference.

6.2. Cosmetic deterioration study

Finkelstein and Wolfe [35] report on a study in which breast cancer patients were treated with either radiation alone ($n_1 = 46$) or radiation plus adjuvant chemotherapy ($n_2 = 48$), and the objective was to compare the two groups on the basis of time to deterioration of the appearance of the breast. In the radiation alone group, 25 observations were right censored and the remaining were interval censored, and there were 13 and 35 right and interval censored observations in the radiation plus chemotherapy group. These data have been analyzed by numerous authors demonstrating methods for analysis of interval censored data; accordingly, although we believe the sample sizes are rather small to support reliable inferences using SNP or any non- or semiparametric approach, we apply these methods for illustration.

The SNP procedure using HQ chose the exponential base density with $K = 0$ in both groups. Figure 3 shows these fits with the NPML estimates superimposed. Using the delta method to

estimate the variance of (10), which is suspect given the small sample sizes, yields a squared SNP test statistic of 8.51 (p-value=0.004). Alternatively, using a bootstrap variance estimate ($B = 100$) leads to a squared SNP test statistic of 8.54 (p-value=0.005). Both are consistent with the chi-square test statistic of Finkelstein and Wolfe [35] of 6.83 (p-value=0.009).

6.3. Massachusetts Health Care Panel Study

Pan and Chappell [1,2] analyze data from the Massachusetts Health Care Panel Study (MHCPs), for which the event time of interest T is the age at which a senior citizen aged 65 or older loses functional independence, defined as inability to carry out activities of daily living. Participants who were all at least 65 years old at baseline completed a survey at study entry as well as up to four subsequent surveys. As only seniors who were functionally independent at baseline were included, all event times were left truncated; moreover, as functional independence was assessed only at survey visits, event times were known only to have occurred between visits or after the last visit, leading to interval and right censoring. Pan and Chappell consider two groups: nonpoor females, with $n_1 = 607$, of which 321 observations were interval censored and the rest were right censored; and nonpoor males, $n_2 = 421$, of which 183 and 238 event times were right and interval censored, respectively.

Denote the observed data as (U_i, L_i, R_i) , $i = 1, \dots, n$, where $U_i \geq 65$ is the truncation time (age at entry) for subject i ; and $L_i \geq U_i$ and $R_i > U_i$ are the endpoints of the interval in which i lost functional independence, with $R_i = \infty$ if the event time is right censored at L_i . Because all participants were 65 or older, analysis is of necessity limited to inference on the conditional survival function $P(T > t \mid T \geq 65)$, $t \geq 65$. We considered two uses of SNP in this context: (i) Let $S(t) = P(T > t \mid T \geq 65)$, which may be operationalized by redefining (U_i, L_i, R_i) to be the original values minus 65; or, (ii) let $S(t) = P(T > t)$, the unconditional survival function, recognizing that while inference on $S(t)$ itself is infeasible, only inference on $S(t)/S(u)$ for $t \geq u \geq 65$ is required (or possible). In both cases, represent the assumed density $f(t)$ of $S(t)$ so defined by SNP; in either, the loglikelihood is then

$$\ell_K(\theta) = \sum_{i=1}^n \log \left\{ \frac{S_K(L_i; \theta) - S_K(R_i; \theta)}{S_K(U_i; \theta)} \right\}. \quad (11)$$

We implemented (i) and (ii) and found that HQ was smaller overall under (ii). We conjecture that approach (ii), which parameterizes $P(T > t \mid T \geq u)$ offers more flexibility than (i), which parameterizes $P(T > t \mid T \geq 65)$ directly. Under (ii), HQ chose $K_1 = 0$ for females and $K_2 = 2$ for males, both with the exponential base density. Figure 3 shows the estimates of the conditional survival functions $P(T > t \mid T \geq 65)$, $S_K(t; \theta)/S_K(65; \theta)$, $t \geq 65$, based on these fits, which are comparable to Figure 2 and figure 3 of Reference [1]. To test whether the true conditional survival functions differ between genders, we substituted these ratios for S_{1,K_1} and S_{2,K_2} in (10) and computed a corresponding standard error estimate using the delta method, with $w(t) \equiv 1$. The SNP test statistic is 3.50; using the bootstrap with $B = 100$, we obtained a statistic of 3.02. Both statistics suggest strong evidence of a difference.

7. DISCUSSION

We have proposed an approach to estimation of the survival function of an event-time random variable based on arbitrarily censored and/or truncated data under the assumption that its distribution has a ‘smooth’ density. The SNP representation strategy we recommend yields ‘smooth’ estimators offering efficiency gains over completely nonparametric methods and performance comparable to or in some cases exceeding that of other smoothing

methods. An SNP-based test statistic to compare two survival functions exhibits reasonable operating characteristics. An advantage of methods based on smoothness assumptions such as the SNP is that arbitrary patterns of censoring and truncation are handled easily in a unified framework. Additionally, the SNP representation provides an informal basis for detecting departures from a proposed parametric model; and, although we did not demonstrate here, because an efficient rejection sampling algorithm for simulation from an SNP density is available [19], once a final SNP fit is determined, any functional of the survival distribution may be approximated by Monte Carlo.

In applications of SNP in other statistical models [20,21,22], which used only a normal base density, the density estimate sometimes exhibits artefactual “bumps” in the tails, particularly when the true density is very skewed. Considering two base density representations (standard normal and exponential) is sufficient to approximate any plausible ‘smooth’ survival function with a small number of ‘parameters;’ i.e., small K . Other formulations may be conceived using other base densities with a moment generating function. The SNP representation can also be adapted to situations where the event time distribution is thought to have finite support by taking the base density to have finite support.

Our simulations, both those here and others not reported, indicate that SNP-based Wald confidence intervals using delta method standard errors treating the chosen K -base density representation as fixed offer satisfactory operating characteristics as long as the information in the data is sufficiently large. We thus recommend this approach only under these conditions; otherwise we recommend use of Wald intervals based on nonparametric bootstrap standard errors. Derivation of formal large sample properties of the SNP survival function estimator is an open problem and beyond the scope of this paper, and would involve determination of the appropriate rate at which K should be chosen with n .

Computationally, the SNP is straightforward and efficient to implement using standard optimization software. SAS macros implementing the proposed procedures are available from the first author upon request.

The SNP representation may be extended straightforwardly to approximate the survival function of correlated bivariate event times ($q = 2$) under arbitrary censoring; we are currently studying this approach and will report on these developments elsewhere.

Acknowledgments

This research was supported by Grants R01 CA085848 from the National Cancer Institute and R37 AI031789 from the National Institute of Allergy and Infectious Diseases.

Contract/grant sponsor: National Institutes of Health; contract/grant number: R01-CA085848,R37-AI031789

REFERENCES

1. Pan W, Chappell R. Computation of the NPMLE of distribution functions for interval censored and truncated data with applications to the Cox model. *Computational Statistics and Data Analysis*. 1998; 28:33–50.
2. Pan W, Chappell R. Estimation in the Cox proportional hazards model with left-truncated and interval-censored data. *Biometrics*. 2002; 58:64–70. [PubMed: 11890328]
3. Hougaard, P. *Analysis of Multivariate Survival Data*. New York: Springer; 2000.
4. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53:457–481.
5. Turnbull BW. The empirical distribution function with arbitrarily grouped, censored, and truncated data. *Journal of the Royal Statistical Society, Series B*. 1976; 38:290–295.

6. Lindsey JC, Ryan LM. Methods for interval censored data. *Statistics in Medicine*. 1998; 17:219–238. [PubMed: 9483730]
7. Huang, J.; Wellner, JA. Interval censored data: A review of recent progress. In: Lin, DY.; Fleming, TR., editors. *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*; New York: Springer; 1997. p. 123–169.
8. Pan W. Smooth estimation of the survival function for interval censored data. *Statistics in Medicine*. 2000; 19:2611–2624. [PubMed: 10986537]
9. Marron S, Padgett WJ. Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *Annals of Statistics*. 1987; 15:1520–1535.
10. Kooperberg C, Stone CJ. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics*. 1992; 1:301–328.
11. Koo JY, Kooperberg C, Park J. Logspline density estimation under censoring or truncation. *Scandinavian Journal of Statistics*. 1999; 26:87–105.
12. Komárek A, Lesaffre E, Hilton JF. Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*. 2005; 14:726–745.
13. Chaubey YP, Sen PK. On smooth estimation of survival and density functions. *Statistics and Decisions*. 1996; 14:1–22.
14. Kim C, Park BU, Kim W, Lim C. Bezier curve smoothing of the Kaplan-Meier estimator. *Annals of the Institute of Statistical Mathematics*. 2003; 55:359–367.
15. Anderson, PK.; Borgan, Ø.; Gill, RD.; Keiding, N. *Statistical Models Based on Counting Processes*. New York: Springer-Verlag; 1993.
16. Keiding N. Age-specific incidence and prevalence: statistical perspective (with discussion). *Journal of the Royal Statistical Society, Series A*. 1991; 154:371–412.
17. Zhang M, Davidian M. “Smooth” semiparametric regression analysis for arbitrarily censored time-to-event data. *Biometrics*. in press.
18. Gallant AR, Nychka. Semi-nonparametric maximum likelihood estimation. *Econometrica*. 1987; 55:363–390.
19. Gallant, AR.; Tauchen, GE. A nonparametric approach to nonlinear time series analysis: Estimation and simulation. In: Brillinger, D.; Caines, P.; Geweke, J.; Parzen, E.; Rosenblatt, M.; Taqqu, MS., editors. *New Directions in Time Series Analysis, Part II*. New York: Springer-Verlag; 1992. p. 71–92.
20. Davidian M, Gallant AR. The nonlinear mixed effects model with a smooth random effects density. *Biometrika*. 1993; 80:475–488.
21. Zhang D, Davidian M. Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics*. 2001; 57:795–802. [PubMed: 11550930]
22. Song X, Davidian M, Tsiatis AA. A semiparametric likelihood approach for joint modeling of longitudinal and time-to-event data. *Biometrics*. 2002; 58:742–753. [PubMed: 12495128]
23. Fenton VM, Gallant AR. Qualitative and asymptotic performance of SNP density estimators. *Journal of Econometrics*. 1996; 74:77–118.
24. Kalbfleisch, JD.; Prentice, RL. *The Statistical Analysis of Failure Time Data*. 2nd edn. New York: Wiley; 2002.
25. SAS Institute Inc. SAS Online Doc 9.1.3. Cary, NC: SAS Institute, Inc; 2006.
26. Kooperberg C, Clarkson DB. Hazard regression with interval-censored data. *Biometrics*. 1997; 53:1485–1494. [PubMed: 9423263]
27. Volinsky CT, Raftery AE. Bayesian information criterion for censored survival models. *Biometrics*. 2000; 56:256–262. [PubMed: 10783804]
28. Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics*. 1989; 45:497–507. [PubMed: 2765634]
29. Petroni GR, Wolfe RA. A two-sample test for stochastic ordering with interval-censored data. *Biometrics*. 1994; 50:77–87. [PubMed: 8086617]
30. Kooperberg, C. The logspline package.
<http://cran.r-project.org/web/packages/logspline/index.html>.

31. Kooperberg, C. The polyspline package.
<http://cran.r-project.org/web/packages/polspline/index.html>.
32. Komárek, A. The smoothSurv package.
<http://cran.r-project.org/web/packages/smoothSurv/index.html>.
33. Hammer SM, Katzenstein DA, Hughes MD, Gundaker H, Schooley RT, Haubrich RH, Henry WK, Lederman MM, Phair JP, Niu M, Hirsch MS, Merigan TC. The AIDS Clinical Trials Group Study 175 Study Team. A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter. *New England Journal of Medicine*. 1996; 335:1081–1089. [PubMed: 8813038]
34. Davidian M, Tsiatis AA, Leon S. Semiparametric estimation of treatment effect in a pretest-posttest study with missing data (with Discussion). *Statistical Science*. 2005; 20:261–301. [PubMed: 19081743]
35. Finkelstein DM, Wolfe RA. A semiparametric model for regression analysis of interval-censored failure time data. *Biometrics*. 1985; 41:933–945. [PubMed: 3830259]

Figure 1. 5th and 95th percentiles (dashed lines) and average (bold dashed line) of 1000 estimates of the true survival function (solid line) based on the SNP, **oldlogspline**, **smoothSurv**, and KM estimators, respectively, for each true distribution and 30% right censoring. Reading down, the rows correspond to the true lognormal, Weibull, exponentiated normal mixture, and Weibull-lognormal mixture distributions, with the SNP (left), **oldlogspline** (left-middle), **smoothSurv** (right-middle), and KM (right) results shown reading across.

Figure 2. 5th and 95th percentiles (dashed lines) and average (bold dashed line) of 1000 estimates of the true survival function (solid line) based on the SNP, **oldlogspline**, **smoothSurv**, and NPMLE estimators, respectively, for each true distribution and interval censoring. Reading down, the rows correspond to the true lognormal, Weibull, exponentiated normal mixture, and Weibull-lognormal mixture distributions, with the SNP (left), **oldlogspline** (left-middle), **smoothSurv** (right-middle), and NPMLE (right) results shown reading across.

Figure 3.

Estimated survival functions for the three applications in Section 6. The leftmost plot shows SNP (smooth curves) and KM (jagged curves) estimates for the ZDV monotherapy (lower) and combined (upper) treatment groups in ACTG 175. The center plot shows the SNP (smooth) and NPML (step function) estimated survival functions for the radiation alone (upper) and radiation plus chemotherapy (lower) groups in the cosmetic deterioration study. The rightmost plot shows SNP estimates of the conditional survival functions for the nonpoor female (upper) and nonpoor male (lower) groups in the MHCPS.

Number of Monte Carlo data sets (out of 1000) for which HQ chose each K -base density representation for the four true survival distributions.

Table 1

Base density	Standard Normal				Standard Exponential			
K	0	1	2	3	0	1	2	3
	30% censoring							
Lognormal	899	54	16	0	0	14	10	7
Weibull	0	0	12	1	872	93	19	3
Exp Norm Mixture	0	2	970	22	0	0	0	6
Weib-Lognorm Mixture	0	0	3	19	0	0	156	822
	50% censoring							
Lognormal	892	49	17	1	2	20	12	7
Weibull	6	1	21	4	838	104	20	6
Exp Norm Mixture	1	498	458	16	0	2	8	17
Weib-Lognorm Mixture	0	0	434	88	0	0	95	383

Table II

Simulation results for 1000 Monte Carlo data sets for estimating $S(t)$ at t corresponding to $S(t) = 0.8, 0.6, 0.4, 0.2$, right censored data, $n = 400$. KM-CP and SNP-CP are Monte Carlo coverage probability of nominal 95% Wald confidence intervals based on the KM and preferred SNP estimator (using Greenwood and delta method standard errors). RE-KM/SNP, RE-LS/SNP, and RE-SS/SNP are Monte Carlo relative efficiency of the SNP estimator to the KM, oldlogpline, and smoothSurv estimators, respectively (defined in text). No results shown for oldlogpline, exponentiated normal mixture data, 50% censoring due to convergence issues. Results not shown for t beyond the range of uncensored data. RE-P is Monte Carlo relative efficiency of SNP relative to fitting a correct parametric lognormal or Weibull model, and P-CP is coverage probability of 95% Wald intervals based on the correct parametric model fit. Standard errors of Monte Carlo coverage ≈ 0.007 .

$S(t)$		30% censoring				50% censoring			
		0.8	0.6	0.4	0.2	0.8	0.6	0.4	0.2
Lognormal	KM-CP	0.943	0.947	0.956	0.947	0.945	0.951	0.944	0.932
	SNP-CP	0.942	0.944	0.953	0.945	0.950	0.950	0.941	0.923
	RE-KM/SNP	1.44	1.41	1.40	1.46	1.44	1.36	1.37	2.15
	RE-LS/SNP	1.54	1.17	1.24	1.24	1.32	1.18	1.13	1.54
	RE-SS/SNP	1.05	1.05	1.02	1.05	1.04	1.02	1.03	0.93
	RE-P	0.97	0.94	0.95	0.98	0.94	0.93	0.97	0.74
Weibull	P-CP	0.946	0.947	0.947	0.947	0.950	0.949	0.943	0.947
	KM-CP	0.945	0.953	0.946	0.947	0.945	0.955	0.946	0.951
	SNP-CP	0.940	0.945	0.948	0.947	0.946	0.946	0.948	0.953
	RE-KM/SNP	1.40	1.44	1.47	1.53	1.38	1.46	1.39	1.50
	RE-LS/SNP	1.23	1.20	1.19	1.14	1.20	1.23	1.14	1.10
	RE-SS/SNP	1.24	1.32	1.30	1.06	1.25	1.30	1.05	1.25
Exp Norm	RE-P	0.95	0.93	0.93	0.96	0.92	0.90	0.95	0.89
	P-CP	0.944	0.953	0.953	0.951	0.948	0.953	0.950	0.951
	KM-CP	0.962	0.952	0.939	0.937	0.957	0.952	0.940	–
	SNP-CP	0.949	0.949	0.943	0.952	0.931	0.942	0.947	–
	RE-KM/SNP	1.31	1.22	1.14	1.04	1.23	1.15	1.03	–
	RE-LS/SNP	1.49	1.17	1.14	1.06	–	–	–	–
Mixture	RE-SS/SNP	1.03	1.04	1.08	1.10	0.96	1.08	1.47	–
	KM-CP	0.944	0.947	0.949	0.939	0.944	0.944	0.938	–
	SNP-CP	0.940	0.946	0.942	0.935	0.946	0.937	0.938	–

		30% censoring				50% censoring			
		0.8	0.6	0.4	0.2	0.8	0.6	0.4	0.2
Mixture	$S(\ell)$								
	RE-KM/SNP	1.26	1.09	1.12	1.42	1.06	1.10	1.21	–
	RE-LS/SNP	1.08	1.00	1.04	1.20	1.10	1.05	1.13	–
	RE-SS/SNP	1.04	1.06	1.02	0.97	1.06	0.92	1.00	–

Simulation results based on 100 Monte Carlo data sets for estimating $S(t)$ at t corresponding to $S(t) = 0.8, 0.6, 0.4, 0.2, 30\%$ right censored data, $n = 100$. KM-CP, SNP-CP, RE-KM/SNP, RE-LS/SNP, and RE-SS/SNP are as in Table II. SNP-CP_{boot}, LS-CP_{boot}, and SS-CP_{boot} are Monte Carlo coverage probabilities for SNP, oldlogpline, and smoothSurv of nominal 95% Wald confidence intervals based on bootstrap standard errors. Standard errors of Monte Carlo coverage entries ≈ 0.02 .

Table III

$S(t)$	0.8	0.6	0.4	0.2	0.8	0.6	0.4	0.2
	Lognormal				Weibull			
KM-CP	0.94	0.95	0.95	0.88	0.96	0.94	0.93	0.91
SNP-CP	0.91	0.92	0.90	0.90	0.96	0.94	0.92	0.91
SNP-CP _{boot}	0.95	0.97	0.97	0.94	0.96	0.96	0.95	0.95
LS-CP _{boot}	0.92	0.96	0.95	0.93	0.97	0.94	0.94	0.94
SS-CP _{boot}	0.93	0.97	0.97	0.98	0.98	0.94	0.94	0.97
	Exp Norm Mixture				Weib-Lognorm Mixture			
KM-CP	0.95	0.94	0.95	0.93	0.94	0.94	0.94	0.95
SNP-CP	0.96	0.94	0.97	0.93	0.91	0.93	0.89	0.93
SNP-CP _{boot}	0.96	0.97	0.96	0.96	0.93	0.96	0.97	0.97
LS-CP _{boot}	0.98	0.95	0.99	0.98	0.96	0.95	0.97	0.97
SS-CP _{boot}	0.98	0.96	0.95	0.91	0.93	0.96	0.94	0.97

Simulation results based on 1000 Monte Carlo data sets for estimating $S(t)$ at t corresponding to $S(t) = 0.8, 0.6, 0.4, 0.2$, interval censored data, $n = 500$. SNP-CP, RE-LS/SNP, and RE-SS/SNP are defined as in Table II, and RE-NP/SNP is Monte Carlo relative efficiency of the SNP estimator to the NPML. Standard errors of Monte Carlo coverage entries ≈ 0.007 .

Table IV

$S(t)$	0.8	0.6	0.4	0.2	0.8	0.6	0.4	0.2
	Lognormal				Weibull			
SNP-CP	0.946	0.945	0.940	0.945	0.944	0.944	0.940	0.947
RE-NP/SNP	4.20	3.16	2.38	2.09	3.77	2.96	2.60	2.08
RE-LS/SNP	1.55	1.28	1.30	1.09	1.12	1.11	1.13	1.05
RE-SS/SNP	1.05	1.03	0.97	0.98	1.23	1.20	1.16	1.06
	Exp Norm Mixture				Weib-Lognorm Mixture			
SNP-CP	0.943	0.941	0.941	0.935	0.963	0.934	0.942	0.940
RE-NP/SNP	4.79	3.82	1.85	1.32	2.01	1.59	1.84	1.96
RE-LS/SNP	3.19	1.85	0.95	0.96	0.85	0.94	0.95	0.98
RE-SS/SNP	1.08	0.95	0.94	0.98	1.01	0.99	0.99	0.94