

Simultaneous Inferences on the Contrast of Two Hazard Functions with Censored Observations

Peter B. Gilbert,^{1,*} L. J. Wei,² Michael R. Kosorok,³ and John D. Clemens⁴

¹Fred Hutchinson Cancer Research Center and Department of Biostatistics,
University of Washington, Seattle, Washington 98109, U.S.A.

²Department of Biostatistics, Harvard School of Public Health,
Boston, Massachusetts 02115, U.S.A.

³Department of Statistics and Biostatistics and Department of Medical Informatics,
University of Wisconsin, Madison, Wisconsin 53792, U.S.A.

⁴National Institute of Child Health and Human Development,
Bethesda, Maryland 20892, U.S.A.

*email: pgilbert@scharp.org

SUMMARY. In survival analysis, often times the pattern of instantaneous risk over time is more interesting than that of the cumulative risk. For this case, a nonparametric hazard function estimate is more appropriate for summarizing the risk experience of a group of patients than the corresponding Kaplan–Meier estimate. In comparing a new treatment with a standard therapy, it is important to know if the treatment loses its potency during the follow-up period, and if it does, one would like to know when it becomes ineffective. Unfortunately, with a plot of the differences of two Kaplan–Meier curves, it is rather difficult to capture such temporal trends. In this article, we propose simple procedures for constructing confidence bands for the contrast of two hazard functions with censored data. The simultaneous interval estimates are quite useful for identifying possible values of the contrast over time with a certain degree of confidence. The new proposals are illustrated with an example and a small simulation study.

KEY WORDS: Confidence bands; Kernel estimation; Martingale; Survival analysis.

1. Introduction

To compare two groups of subjects with a time-to-event endpoint, the Kaplan–Meier curves (Kaplan and Meier, 1958) are routinely used to summarize the group difference graphically. For example, in a placebo-controlled, three-arm vaccine trial conducted in rural Bangladesh among children and adult women, two similar oral vaccines were evaluated for preventing cholera disease (Clemens et al., 1990). For this trial, 62,178 children and women were randomly assigned to three groups, of whom 41,342 were vaccinated and 20,836 received placebo. The response variable for the study was the time between 14 days after the third immunization and the first diarrheal episode in which *Vibrio cholerae* 01 was isolated. During the 3-year follow-up period, 258 vaccine recipients and 266 placebo recipients experienced the outcome. Figure 1a presents the Kaplan–Meier curves for the pooled vaccine group and the placebo group, and Figure 1b gives the differences of these two curves. To show the degree of uncertainty of the estimated differences over the entire time span of interest, one may use the simulation technique proposed by Parzen, Wei, and Ying (1997) to construct, e.g., 95% confidence bands for the difference of the two survival functions (see the dashed lines in Figure 1b). It is rather difficult, however, to visualize

the temporal effect of the vaccine with such plots based on cumulative incidence rates (Efron, 1988). For instance, from Figure 1b, it appears that the vaccine was effective in preventing the aforementioned diarrheal episode, but it is not clear whether the vaccine lost its potency during the follow-up period. Moreover, if it did, it is important to know when and how fast the vaccine became ineffective.

Figure 2 gives the corresponding hazard functions in terms of their nonparametric kernel estimates (see Section 2 for details). Note that the hazard estimates change markedly over time, with a similar pattern for the two groups, indicating seasonal and annual variations in cholera incidence. This observation is consistent with the known epidemiology of cholera in Bangladesh (Clemens et al., 1990) but is difficult to detect from the Kaplan–Meier plots in Figure 1. To see the differences of two hazard functions graphically, one may plot the logarithm of the ratio of the two hazard function estimates over time (see Figure 3). We now see that the vaccine efficacy gradually diminished, especially after 2 years from immunization. This type of plot contains more information about the temporal effects of the vaccines than the plots in Figure 1. To draw inferences about a contrast of the two underlying hazard functions over a preselected time interval, pictorially

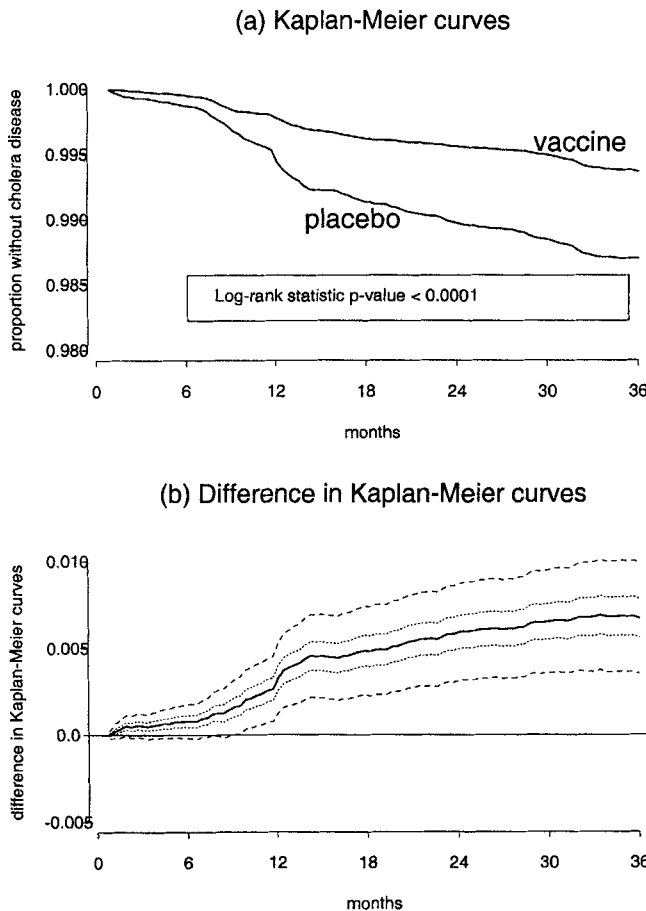


Figure 1. a. Kaplan-Meier estimates of survival curves for the placebo and vaccine groups in the cholera vaccine trial. b. Point estimates (solid curves) and 95% pointwise (dotted curves) and simultaneous (dashed curves) confidence intervals for the differences of the two survival curves.

simultaneous confidence interval estimates are quite informative because they provide possible values of this contrast over the entire time span of interest with a certain degree of confidence. It is important to note that, if one views the contrast of two hazard rates as a function of time, confidence bands are more appropriate than their pointwise counterparts for quantifying the degree of uncertainty of the point estimates. To the best of our knowledge, there are no methods available for constructing such confidence bands in the literature.

In this article, we propose simulation-based and analytical procedures for obtaining simultaneous confidence intervals for a contrast of two hazard functions. Sections 2 and 3 develop the procedures in two steps, the first for obtaining confidence bands for a single hazard function and the second for a contrast of two hazard functions. The methods are illustrated with the above vaccine trial example and a small simulation experiment in Section 4.

Our approach for constructing confidence bands via kernel smoothing has been taken by many others for constructing confidence bands for densities and in nonparametric regression. This literature includes Hall and Titterton (1988), Hall and Owen (1993), Härdle and Bowman (1988), Eubank

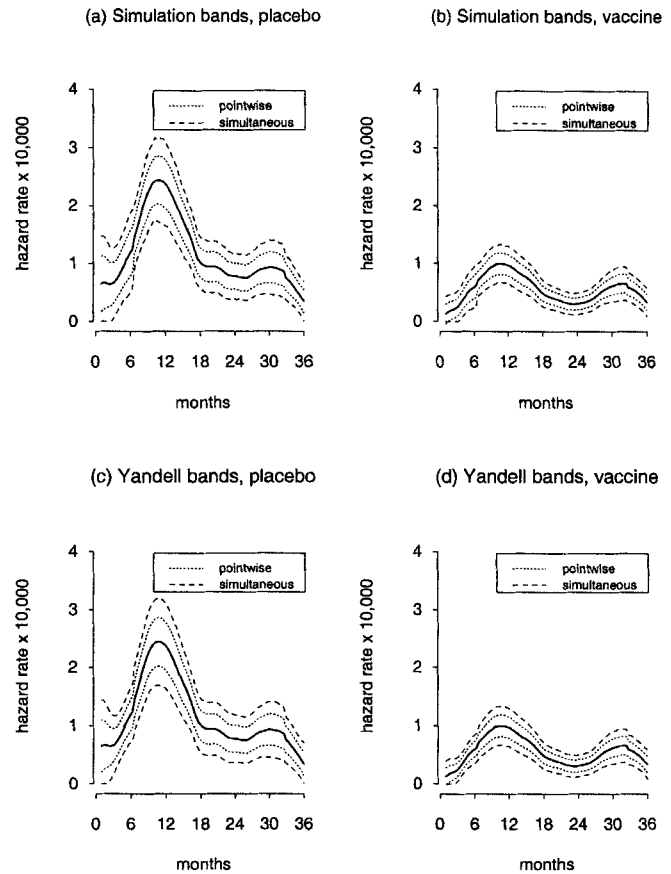


Figure 2. Kernel estimates (solid curves) and 95% pointwise (dotted curves) and simultaneous (dashed curves) confidence intervals for the hazard functions of the placebo and vaccine groups in the cholera vaccine trial.

and Speckman (1993), and Xia (1998). To our knowledge, the literature is restricted to the one-sample problem; thus, the present work is novel within a broad area for its focus on the two-sample problem.

2. Confidence Bands for a Hazard Function

2.1 Nonparametric Hazard Function Estimator

Let X_{ij} be the minimum of the failure time and censoring time for the j th subject in the i th group, $j = 1, \dots, n_i$; $i = 1, 2$. Let Δ_{ij} be the censoring indicator, equal to one if the failure time is observed and zero otherwise. The data consist of two independent samples $\{(X_{ij}, \Delta_{ij}), j = 1, \dots, n_i\}$, $i = 1, 2$. Within each group, we assume that the failure time and censoring time are independent. Let $\lambda_i(t)$ be the hazard function for subjects in the i th group and let $[t_1, t_2]$ be a fixed time interval such that $\text{pr}(X_{i1} < t_1, \Delta_{i1} = 1) > 0$, and $\text{pr}(X_{i1} > t_2, \Delta_{i1} = 1) > 0$, $i = 1, 2$. We consider inference on each $\lambda_i(t)$, $i = 1, 2$, in a preselected time interval $[u_1, u_2]$ contained in $[t_1, t_2]$.

For $t \in [u_1, u_2]$, $\lambda_i(t)$ can be consistently estimated by a kernel estimator,

$$\hat{\lambda}_i(t) = \frac{1}{b_i} \int_{t_1}^{t_2} K\left(\frac{t-s}{b_i}\right) d\hat{\Lambda}_i(s), \quad (1)$$

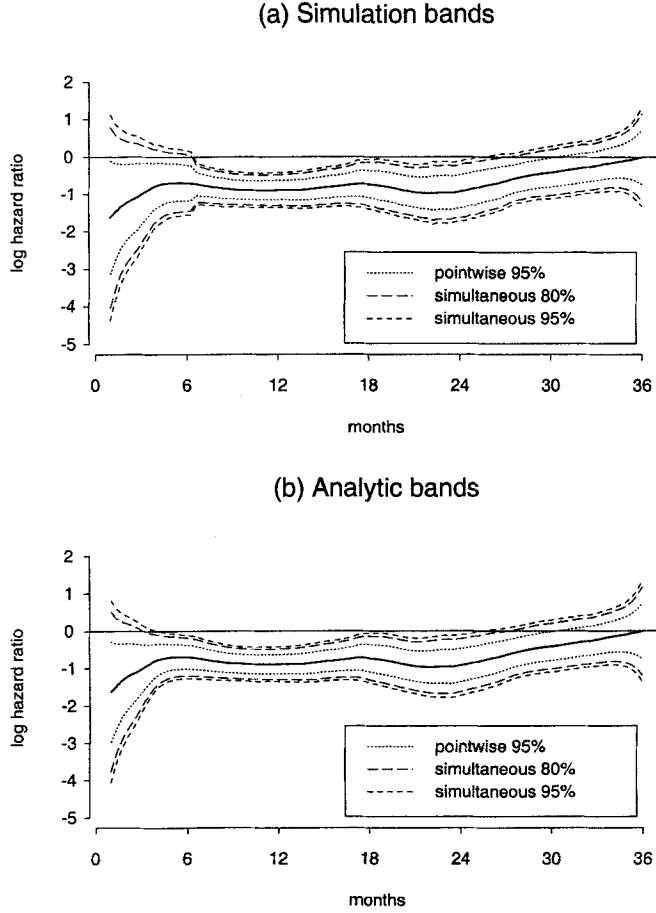


Figure 3. Point estimate (solid curves), 95% pointwise (dotted curves), 80% simultaneous (hatched curves), and 95% simultaneous (dashed curves) confidence intervals for the log hazard ratio (vaccine/placebo) in the cholera vaccine trial.

where $\hat{\Lambda}_i(t) = \int_0^t (1/Y_i(s)) dN_i(s)$ is Nelson (1972) and Aalen's (1978) estimator of the cumulative hazard function $\Lambda_i(t) = \int_0^t \lambda_i(s) ds$, with $N_i(t) = \sum_{j=1}^{n_i} N_{ij}(t) = \sum_{j=1}^{n_i} I(X_{ij} \leq t, \Delta_{ij} = 1)$, $Y_i(t) = \sum_{j=1}^{n_i} Y_{ij}(t) = \sum_{j=1}^{n_i} I(X_{ij} \geq t)$, and $I(\cdot)$ the indicator function. The kernel function K is a bounded function with support $[-1, 1]$, integral one, mean zero, and positive second moment. The bandwidth b_i is a positive quantity that indicates the window $[t - b_i, t + b_i]$ over which the Nelson-Aalen estimator is smoothed. The kernel estimator has been studied by Rice and Rosenblatt (1976), Yandell (1983), Ramlau-Hansen (1983), and Tanner and Wong (1983), among others. Because the kernel estimate $\hat{\lambda}_i(t)$ of (1) is valid only for $t \in [t_1 + b_i, t_2 - b_i]$, we use the method of Gasser and Müller (1979) to extend the estimate to the tail regions. Specifically, for t in the lower tail region $[u_1, t_1 + b_i]$, we estimate $\lambda_i(t)$ by (1) with kernel function $K(\cdot)$ replaced by $K_t(x) = K(x)(\gamma_t + \psi_t x)$, where $x \in [-1, (t - t_1)/b_i]$ and the scalars γ_t and ψ_t are determined by the conditions $\int_{-1}^{(t-t_1)/b_i} K_t(x) dx = 1$ and $\int_{-1}^{(t-t_1)/b_i} x K_t(x) dx = 0$. For t in the upper tail region $(t_2 - b_i, u_2]$, the symmetric counterpart is used.

2.2 Confidence Bands

To approximate the distribution of $\hat{\lambda}_i(t)$, the choice of bandwidth plays an important role. For example, if we choose b_i such that $n_i b_i \rightarrow \infty$ and $n_i^{1/3} b_i \rightarrow 0$ as $n_i \rightarrow \infty$, then for a fixed t ,

$$\begin{aligned} \hat{\lambda}_i(t) - \lambda_i(t) &\approx U_i(t) \\ &= \frac{1}{b_i} \int_{t_1}^{t_2} K\left(\frac{t-s}{b_i}\right) J_i(s) (Y_i(s))^{-1} \\ &\quad \times d \left\{ \sum_{j=1}^{n_i} M_{ij}(s) \right\}, \quad u_1 \leq t \leq u_2, \end{aligned} \quad (2)$$

$J_i(t) = I(Y_i(t) > 0)$, $M_{ij}(t) = N_{ij}(t) - \int_0^t \lambda_i(s) Y_{ij}(s) ds$, and $n_i^{1/2} b_i^{1/2} U_i(t)$ is asymptotically normal (see Andersen et al., 1993, Theorem IV.2.4).

As suggested in the literature for nonparametric function estimation, the bandwidth can alternatively be chosen in a less *ad hoc* way to satisfy an optimality criterion, e.g., to minimize an asymptotic approximation to the mean integrated squared error $\text{MISE}(\hat{\lambda}_i(\cdot)) = E \int [\hat{\lambda}_i(t) - \lambda_i(t)]^2 dt$ (see Andersen et al., 1993, p. 240, formulas (4.2.25) and (4.2.26)). The optimizing bandwidth b_i satisfies $n_i^{1/5} b_i \rightarrow k_i \in (0, \infty)$, and with this choice, $\hat{\lambda}_i(t) - \lambda_i(t) - \widehat{\text{bias}}(\hat{\lambda}_i(t)) \approx U_i(t)$, where $U_i(t)$ is given in (2) and $n_i^{2/5} U_i(t)$ is asymptotically normal (see Andersen et al., 1993, Theorem IV.2.5). It is important to note that this approximation requires a bias correction term $\widehat{\text{bias}}(\hat{\lambda}_i(t))$, which can be taken to be an estimate of $(b_i^2/2) \lambda_i''(t) \int_{-1}^1 x^2 K(x) dx$. Here $\lambda_i''(t)$ is the second derivative of $\lambda_i(t)$, which can be estimated by the kernel estimator proposed by Andersen et al. (1993, p. 249) using a bandwidth b_{i1} that converges to zero more slowly than $n_i^{-1/6}$. For nonparametric regression, Eubank and Speckman (1993) recommended choosing $b_{i1} = b_i^{7/5}$, while Härdle and Bowman recommended $b_{i1} = 2b_i$. Here we select b_{i1} to minimize a bootstrap estimate of $\text{MISE}(\hat{\lambda}_i''(\cdot))$ and then select b_i to minimize the approximation of $\text{MISE}(\hat{\lambda}_i(\cdot))$ computed using formula (4.2.26) in Andersen et al. (1993). The theoretical justification for the properties of the above function estimates with data-dependent bandwidths is given in Gilbert and Kosorok (2001).

A bandwidth b_i satisfying $n_i^{1/3} b_i \rightarrow 0$ is smaller than optimal, i.e., it undersmooths, with the advantage that the inherent bias in $\hat{\lambda}_i(t)$ is negligible in the limit. Hall and Owen (1993) and Eubank and Speckman (1993) discussed undersmoothed kernel estimation for densities and in nonparametric regression, respectively.

To approximate the distribution of $U_i(t)$, note that the mean of $M_{ij}(t)$ is zero and its variance is the expected value of $N_{ij}(t)$. Using the random Gaussian multipliers technique proposed by Lin, Fleming, and Wei (1994) and Parzen et al. (1997) for approximating the distribution of a sum of stochastic integrals with respect to $M_{ij}(t)$, we replace $M_{ij}(t)$ in $U_i(t)$ with $N_{ij}(t) Z_{ij}$ and then replace X_{ij} , Δ_{ij} , and $Y_i(t)$ with their observed values x_{ij} , δ_{ij} , and $y_i(t)$, where $\{Z_{ij}, j = 1, \dots, n_i\}$ is a random sample from the standard normal dis-

tribution. This results in a process

$$\tilde{U}_i(t) = \frac{1}{b_i} \sum_{j=1}^{n_i} K\left(\frac{t - x_{ij}}{b_i}\right) I(t_1 \leq x_{ij} \leq t_2) \times \{y_i(x_{ij})\}^{-1} I(y_i(x_{ij}) > 0) \delta_{ij} Z_{ij}. \quad (3)$$

Note that the only stochastic components in $\tilde{U}_i(t)$ are the $\{Z_{ij}\}$. Now, conditional on the data, the mean of $\tilde{U}_i(t)$ is zero and its covariance function evaluated at two time points t and $\tilde{t}$ is

$$b_i^{-2} \int_{t_1}^{t_2} K\left(\frac{t-s}{b_i}\right) K\left(\frac{\tilde{t}-s}{b_i}\right) J_i(s) (Y_i(s))^{-2} d \left[\sum_{j=1}^{n_i} N_{ij}(s) \right].$$

The unconditional expected value of this covariance is precisely the covariance of $U_i(t)$ and $U_i(\tilde{t})$. This implies that, for any given t , $n_i^{2/5} U_i(t)$ and $n_i^{2/5} \tilde{U}_i(t)$ have the same limiting distribution. Unfortunately, because b_i converges to zero, the limiting value of the covariance function is zero whenever $t \neq \tilde{t}$. By Lemma 1.5.9 of Van der Vaart and Wellner (1996), this implies that neither $U_i(\cdot)$ nor $\tilde{U}_i(\cdot)$ have tight limit laws, and thus we cannot use weak convergence plus the continuous mapping theorem applied to the supremum functional in constructing confidence bands as done in Lin et al. (1994) and Parzen et al. (1997). On the other hand, it follows from Yandell (1983) and Bickel and Rosenblatt (1973) that the distribution of the supremum of $U_i(\cdot)$ (suitably standardized and centered) can be approximated by that of the supremum of $\tilde{U}_i(\cdot)$. This follows because standardized and centered versions of the suprema of $U_i(\cdot)$ and of $\tilde{U}_i(\cdot)$ conditional on the data have the same limiting distribution (an extreme value distribution) for a bandwidth b_i satisfying $n_i^{1/5} b_i \rightarrow k_i \in (0, \infty)$. Specifically, for every $x \in (-\infty, \infty)$, with u_{2n_i} a consistent estimator for u_2 as described in Gilbert and Kosorok (2001), $\text{pr}[r_i \{\sup_{u_1 \leq t \leq u_{2n_i}} |U_i(t)/\hat{v}_i(t)| - (r_i + r_i^{-1} \log[\int_{-1}^1 (K'(s))^2 ds / \int_{-1}^1 (K(s))^2 ds]^{1/2} / (2\pi)]\} < x] \rightarrow e^{-2e^{-x}}$, where $r_i = [2 \log[(u_{2n_i} - u_1)/b_i]]^{1/2}$ and $\hat{v}_i(t)$ is the estimated standard deviation of $U_i(t)$, and likewise for $\tilde{U}_i(t)$ standardized and centered in the same way. A proof of this result is given in Gilbert and Kosorok (2001). The proof requires twice continuous differentiability of $\lambda_i(t)$ on $[t_1, t_2]$, which we assume henceforth for $i = 1, 2$.

In practice, one can easily approximate the distribution of $\tilde{U}_i(t)$ by generating a large number, say L , of random samples $\{Z_{ij}\}$. For each realized sample $\{z_{ij}\}$, we calculate a realization of $\tilde{U}_i(t)$. Then the distribution of the supremum of $\tilde{U}_i(\cdot)$, and therefore of $U_i(\cdot)$, can be approximated by the empirical one based on the L sets of realizations of $\tilde{U}_i(\cdot)$. For example, to obtain $1 - \alpha$ -level confidence bands for $\lambda_i(\cdot)$ on an interval $[u_1, u_2] \subseteq [t_1, t_2]$ based on the statistic $\sup_{u_1 \leq t \leq u_2} |U_i(t)/\hat{v}_i(t)|$, one may use the above simulation technique to obtain the cut-off point $c_i(\alpha)$ that satisfies $\text{pr}(\sup_{u_1 \leq t \leq u_2} |\tilde{U}_i(t)/\hat{v}_i(t)| < c_i(\alpha)) \approx 1 - \alpha$. Here, $\hat{v}_i(t)$ is the observed value of $\hat{v}_i(t)$, which can be obtained empirically using the L realizations of $\tilde{U}_i(t)$ or using the variance estimator

$$\hat{v}_i(t)^2 = \widehat{\text{var}}(\hat{\lambda}_i(t))$$

$$= b_i^{-2} \int_{t_1}^{t_2} J_i(s) \{K((t-s)/b_i)/Y_i(s)\}^2 dN_i(s), \quad (4)$$

which is derived from the standard martingale theory for stochastic integrals (cf., Andersen et al., 1993, p. 232). The corresponding confidence bands for $\lambda_i(\cdot)$ on $[u_1, u_2]$ are $[\hat{\lambda}_i(t) - \widehat{\text{bias}}(\hat{\lambda}_i(t)) \pm c_i(\alpha) \hat{v}_i(t) : u_1 \leq t \leq u_2]$. The cut-off point $c_i(\alpha)$ can also be obtained through an analytic approximation to the distribution of $\sup_{u_1 \leq t \leq u_2} |(\hat{\lambda}_i(t) - \lambda_i(t) - \widehat{\text{bias}}(\hat{\lambda}_i(t)))/\hat{v}_i(t)|$ proposed in Theorems 4.1 and 4.2 of Yandell (1983). Because the analytic approximation is less accurate than the simulation-based approximation, the proposed simulation-based bands are expected to be more reliable (see Gilbert and Kosorok, 2001, for details). The improved approximation is akin to Hall's (1991) result that the bootstrap provides a better approximation to the distribution of the supremum of a standardized kernel density estimator than the associated limiting extreme value distribution. Splus/R functions implementing the methods are available from the first author.

3. Confidence Bands for a Contrast of Two Hazard Functions

Let $\eta(\cdot) = g(\lambda_1(\cdot), \lambda_2(\cdot))$ quantify the "difference" between the two hazard functions. For example, $\eta(\cdot) = \log[\lambda_2(\cdot) \div \lambda_1(\cdot)]$ or $\lambda_2(\cdot) - \lambda_1(\cdot)$. For $t \in [u_1, u_2]$, a consistent estimator $\hat{\eta}(t)$ for $\eta(t)$ is $g(\hat{\lambda}_1(t) - \widehat{\text{bias}}(\hat{\lambda}_1(t)), \hat{\lambda}_2(t) - \widehat{\text{bias}}(\hat{\lambda}_2(t)))$. To obtain confidence bands for $\eta(\cdot)$, consider the process $V(t) = (\hat{\eta}(t) - \eta(t))/\hat{v}(t)$, $u_1 \leq t \leq u_2$, where $n = n_1 + n_2$ and $\hat{v}(\cdot)$ is a weight function that converges in probability, uniformly on $[u_1, u_2]$, to a deterministic function $v(\cdot)$. A natural choice of $\hat{v}(t)$ is the estimated standard deviation of $\hat{\eta}(t)$. If $g(\cdot)$ has continuous first partial derivatives, then $[V(t) : u_1 \leq t \leq u_2]$ is approximately equal to the process

$$\hat{v}(t)^{-1} \sum_{i=1}^2 g_i(\lambda_1(t), \lambda_2(t)) \left(\hat{\lambda}_i(t) - \lambda_i(t) - \widehat{\text{bias}}(\hat{\lambda}_i(t)) \right), \quad u_1 \leq t \leq u_2,$$

where g_i is the partial derivative of g with respect to the i th argument, $i = 1, 2$. It follows from (2) that the above process is approximately equal to

$$\hat{V}(t) = \hat{v}(t)^{-1} \sum_{i=1}^2 g_i \left(\hat{\lambda}_1(t) - \widehat{\text{bias}}(\hat{\lambda}_1(t)), \hat{\lambda}_2(t) - \widehat{\text{bias}}(\hat{\lambda}_2(t)) \right) \times U_i(t).$$

When $\hat{v}(t)$ is consistent for the standard deviation of $\hat{\eta}(t)$, then standardized versions of the suprema of $|V(t)|$ and $|\hat{V}(t)|$ have extreme value limiting distributions in the special case in which $b_1 = b_2$ (Gilbert and Kosorok, 2001). In this case, an $1 - \alpha$ -level confidence band for $\eta(\cdot)$ on an interval $[u_1, u_2] \subset [t_1, t_2]$ can be constructed as

$$\{\hat{\eta}(t) \pm k(\alpha) \hat{v}(t) : u_1 \leq t \leq u_2\}, \quad (5)$$

with $k(\alpha) = (d + x/r)$ and $x = \log[-2/\log(1 - \alpha)]$, where $r = (2 \log[(u_2 - u_1)/b_1])^{1/2}$, $d = r + r^{-1} \log[\int_{-1}^1 (K'(s))^2 ds / \int_{-1}^1 (K(s))^2 ds]^{1/2} / (2\pi)$, and $\hat{v}(t)$ is calculated using formula (4) and the delta method.

When $\hat{v}(t)$ is not the estimated standard deviation of $\hat{\eta}(t)$ or when $b_1 \neq b_2$, though, it is difficult if not impossible to obtain the asymptotic distribution of the supremum of $|\hat{V}(\cdot)|$ analytically. On the other hand, in general, the distribution of $\hat{V}(t)$ can be approximated by that of

$$\tilde{V}(t) = \tilde{v}(t)^{-1} \sum_{i=1}^2 g_i \left(\hat{\lambda}_1(t) - \widehat{\text{bias}}(\hat{\lambda}_1(t)), \hat{\lambda}_2(t) - \widehat{\text{bias}}(\hat{\lambda}_2(t)) \right) \times \tilde{U}_i(t),$$

where $\tilde{v}(t)$ is the observed value of $\hat{v}(t)$. In particular, as proved in Gilbert and Kosorok (2001), for optimal bandwidths b_i satisfying $n_i^{1/5} b_i \rightarrow k_i \in (0, \infty)$, $i = 1, 2$, a suitably standardized version of the supremum of $\tilde{V}(\cdot)$ conditional on the data converges to the same distribution (similar to the extreme value distribution) as the standardized version of the supremum of $\hat{V}(\cdot)$. Thus, the distribution of $G = \sup_{u_1 \leq t \leq u_2} |\hat{V}(t)|$ can be approximated by that of $\tilde{G} = \sup_{u_1 \leq t \leq u_2} |\tilde{V}(t)|$. An $1 - \alpha$ -level confidence band for $[\eta(t) : u_1 \leq t \leq u_2]$ is then given by

$$\{\hat{\eta}(t) \pm d(\alpha)\tilde{v}(t) : u_1 \leq t \leq u_2\}, \quad (6)$$

where $d(\alpha)$ is defined by $\text{pr}(\tilde{G} > d(\alpha)) = \alpha$ and can be approximated using, say, L realizations of $\tilde{G}$. If the weight function $\hat{v}(t)$ is the estimated standard deviation of $\hat{\eta}(t)$, it may be calculated empirically using the L realizations of each $\tilde{U}_i(\cdot)$, $i = 1, 2$, or using equation (4) and the delta method.

In the special case that the analytical bands can be computed, the simulation-based bands are expected to behave more reliably because $\tilde{G}$ converges to G more quickly than the closed-form limiting approximation (Gilbert and Kosorok, 2001). Note that, if undersmoothed bandwidths are used, the above procedures yield confidence bands without requiring a bias correction, i.e., with $\widehat{\text{bias}}(\hat{\lambda}_i(t))$ taken to be zero.

4. Example and Simulation Experiment

For the vaccine trial in Bangladesh, let $\lambda_1(\cdot)(\lambda_2(\cdot))$ be the hazard function for trial participants in the placebo (pooled vaccine) group. Let the kernel function be Epanechnikov's kernel, $K(x) = 0.75(1 - x^2)I(|x| \leq 1)$, and let $t_1 = u_1 = 1$ month and $u_2 = 36, t_2 = 38$ months. The interval $[1, 36]$ within $[1, 38]$ was chosen because some events occurred in each group before 1 month and after 38 months. The optimal bandwidths b_{11} and b_i were calculated as described in Section 2.2, giving $b_{11} = 11.62, b_1 = 5.56, b_{12} = 11.64, b_2 = 5.59$. Gasser and Müller's (1979) procedure was used to estimate $\lambda_i(t)$ in the tail regions $[1, 1 + b_i)$ and $(38 - b_i, 36]$.

Figure 2a and 2b shows 95% bias-corrected confidence bands for $\lambda_1(\cdot)$ and $\lambda_2(\cdot)$ over the time interval 1–36 months based on 1000 realizations of $\tilde{U}_1(\cdot)$ and $\tilde{U}_2(\cdot)$, respectively, using empirical standard error estimates $\tilde{v}_i(t)$. The critical values are approximated as $c_1(0.05) = 3.40$ and $c_2(0.05) = 3.49$. For comparison, we also constructed the confidence bands for $\lambda_i(\cdot)$ using Yandell's (1983) second-order approximation to the asymptotic distribution of the supremum of the standardized $\hat{\lambda}_i(\cdot)$ and the variance estimates in (4) (see Figure 2c and 2d). Yandell's bands are very similar to ours, with critical values 3.48 and 3.52 for the placebo and vaccine groups, respectively.

Next, set $\eta(t) = \log\{\lambda_2(t)/\lambda_1(t)\}$. Figure 3a depicts 80 and 95% confidence bands for $\eta(\cdot)$ based on 1000 realizations of $\tilde{G}$. Here the weight function $\hat{v}(\cdot)$ was the empirical standard deviation of $\hat{\eta}(\cdot)$ estimated using 1000 realizations of each $\tilde{U}_i(\cdot)$. The critical values are $d(0.20) = 3.13$ and $d(0.05) = 3.56$. For vaccine trials in general, the vaccine efficacy at time t may be defined as $1 - \lambda_2(t)/\lambda_1(t)$, which has a useful interpretation (Halloran, Haber, and Longini, 1992). For the present case, the estimated vaccine efficacy ($= 1 - e^{\hat{\eta}(t)}$) is highest 6 weeks after the third immunization ($t = 1$ month) at $1 - e^{-1.7} = 0.82$, drops linearly over the next 5 months to about $1 - e^{-0.6} = 0.45$, vacillates between 0.45 and $1 - e^{-1.0} = 0.63$ over the next 18 months, and then wanes steadily to zero during the third year of follow-up. Because the confidence bands are relatively wide during the first 4 months, the inference of relatively high vaccine efficacy during this early period must be interpreted cautiously. Note that the upper confidence band rises above zero at 2 years of follow-up.

Figure 3b shows the analysis based on the analytical confidence bands of (5) using common bandwidth $b = (b_1 + b_2)/2 = 5.575$ and $\hat{v}(t)$ calculated using (4) and the delta method. The confidence bands are very similar to the simulation-based bands, with practically identical 0.20- and 0.05-level critical values 3.12 and 3.56, respectively. We also recalculated the simulation-based confidence bands without the bias correction and again with undersmoothed bandwidths b_i satisfying $n_i^{1/3} b_i \rightarrow k \in (0, \infty)$ and obtained very similar results. Note that the 80% confidence bands are not much narrower than the 95% confidence bands. Eubank and Speckman (1993) observed the same property for their bands in nonparametric regression.

From Figure 3, we may conclude that the vaccines were moderately efficacious in preventing cholera disease during the first 2 years after the vaccination series and then steadily lost their effectiveness, with no efficacy beyond 3 years. This suggests that immunogens that elicit more durable immune responses need to be developed in reformulated cholera vaccines.

A simulation experiment was conducted to evaluate the finite-sample accuracy of $\tilde{U}_i(t)$ as an approximation for $h_i(t) = \hat{\lambda}_i(t) - \lambda_i(t) - \widehat{\text{bias}}(\hat{\lambda}_i(t))$ and of $\tilde{G}$ for $\sup_{u_1 \leq t \leq u_2} |\hat{V}(t)|$. For groups 1 and 2, we repeatedly generated exponential ($\lambda_1 = 0.064$) and Weibull ($\lambda_2 = 0.064$, shape = 1.5) failure times, respectively, with $n_1 = n_2 = 100$ and 25% random censoring in each group. For each of 1000 simulated datasets, we calculated $n_i^{2/5} h_i(18)$, $V(18)$, $\sup_{4 \leq t \leq 32} |n_i^{2/5} h_i(t)|$, G , and their approximations $n_i^{2/5} \tilde{U}_i(18)$, $\tilde{V}(18)$, $\sup_{4 \leq t \leq 32} |n_i^{2/5} \tilde{U}_i(t)|$, $\tilde{G}$, $i = 1, 2$. The empirical distribution functions of these quantities show that the approximations are quite accurate (see Figure 4), although, as expected, approximations at a fixed time point ($t = 18$) are better than approximations of suprema. In additional simulations, we found that the accuracy of the approximations is insensitive to the use of bias adjustment and to the method of computing the standard deviation $\tilde{v}(t)$, but the accuracy of $\sup_{4 \leq t \leq 32} |n_i^{2/5} \tilde{U}_i(t)|$ diminishes for larger bandwidths.

5. Remarks

In a comparative study with respect to a subject's "survival," quantitative summaries of a contrast of two hazard functions

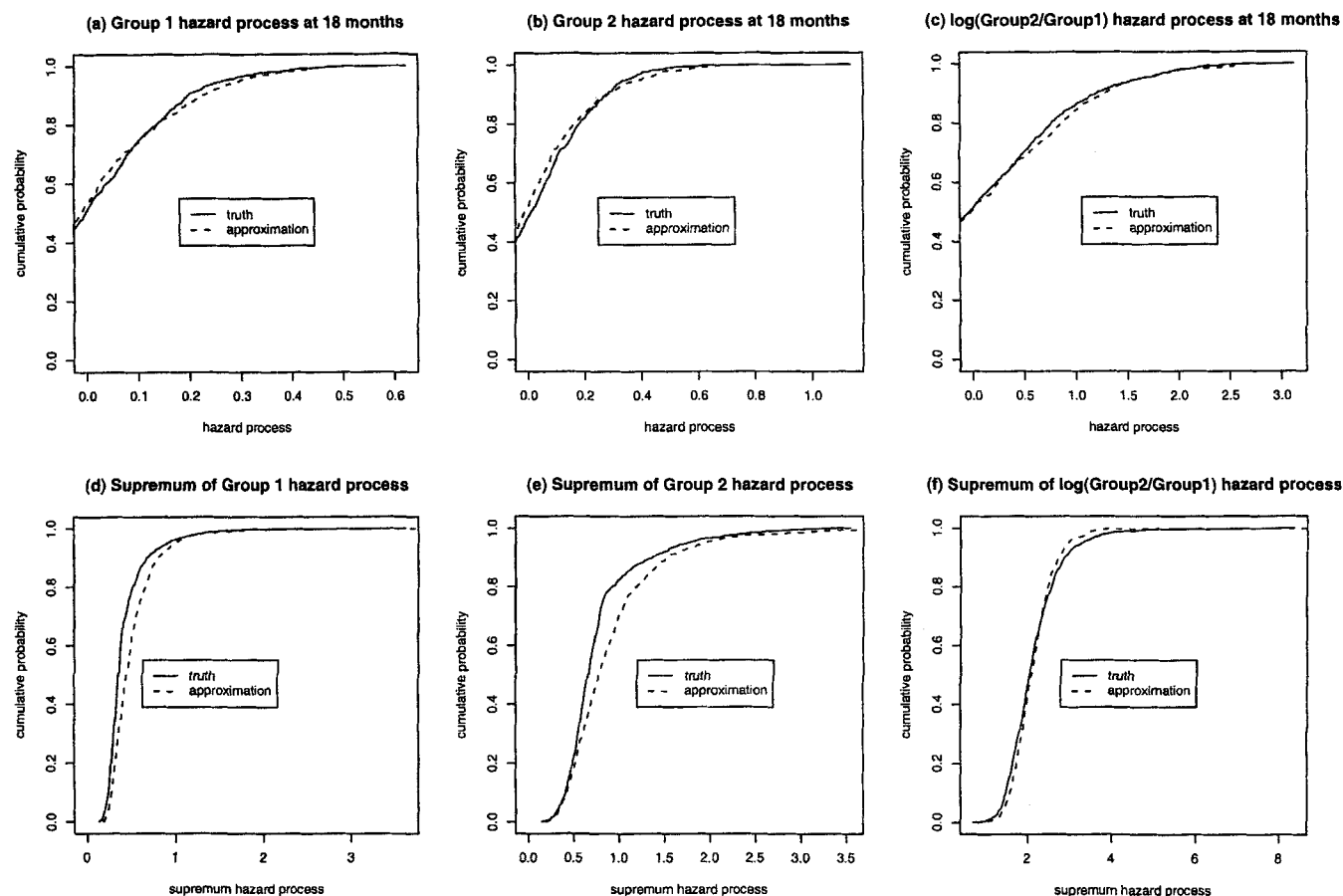


Figure 4. For the simulation study, (a) and (b) show the empirical cumulative distribution function (e.c.d.f.) of $n_i^{2/5}(\hat{\lambda}_i(18) - \lambda_i(18) - \text{bias}(\hat{\lambda}_i(18)))$ and of the approximation $n_i^{2/5}\tilde{U}_i(18)$ for $i = 1, 2$, respectively, (c) shows the e.c.d.f. of $V(18)$ and of the approximation $\tilde{V}(18)$, (d) and (e) show the e.c.d.f. of $\sup_{4 \leq t \leq 32} |n_i^{2/5}(\hat{\lambda}_i(t) - \lambda_i(t) - \text{bias}(\hat{\lambda}_i(t)))|$ and of the approximation $\sup_{4 \leq t \leq 32} |n_i^{2/5}\tilde{U}_i(t)|$ for $i = 1, 2$, respectively, and (f) shows the e.c.d.f. of $\sup_{4 \leq t \leq 32} |V(t)|$ and of the approximation $\tilde{G}$.

can be quite informative. In practice, we highly recommend plotting the point and interval estimates of such contrasts in addition to the routine Kaplan–Meier estimates. Like other measures of the treatment difference over the course of a longitudinal study, the clinical interpretation of the resulting graphical display based on the hazard functions should be interpreted cautiously. The individuals in the risk sets over time are not comparable in the two groups even in a randomized trial setting. On the other hand, the hazard ratio at a given time t postvaccination can be a particularly useful measure of efficacy in a placebo-controlled HIV vaccine trial since randomization and blinding imply that it can be interpreted as the vaccine’s multiplicative reduction in the transmission probability given a single exposure at time t . This mechanism of efficacy is plausible because vaccination may decrease the probability that a given exposing HIV infects a host cell, leading to a proportionate reduction in the probability of established infection.

This article develops simulation-based and analytical procedures for constructing confidence bands about a smooth contrast of two hazard functions. In general, the simulation-

based approach is expected to match or outperform the analytical approach. Bootstrap procedures offer alternatives to our simulation technique. For example, a smooth bootstrap could proceed by estimating the failure time and censoring densities via nonparametric kernel smoothing (using a large bandwidth to oversmooth), independently resampling from the two density estimates for each group to form bootstrap datasets, and for each dataset, calculating the supremum of $V^*(t) = (\hat{\eta}^*(t) - \hat{\eta}(t))/\hat{v}^*(t)$. The result of Hall (1991) for the nonparametric bootstrap in the uncensored case suggests that the smooth bootstrap would have comparable theoretical coverage properties as our simulation-based approach, though this remains to be investigated. The bootstrap would be more computationally intensive than the simulation technique used here.

The bias of kernel hazard estimators and the selection of the bandwidths pose challenges to constructing the confidence bands. We have suggested two approaches, one that bias corrects by kernel estimation of the second derivative of the hazard functions and a simpler approach that undersmooths and does not bias correct. The first method is optimal theoretic-

cally but may suffer if the second derivative is poorly estimated (see the Discussion in Härdle and Bowman, 1988). For applications in which the bias of each hazard function is in the same direction, the bias of the contrast of hazard functions will tend to be diminished. Thus, the bias problem may often be less severe for the two-sample problem than for the one-sample problem.

Note that the confidence bands procedures can be adapted for testing the null hypothesis $H_0: \lambda_1(t) = \lambda_2(t)$ for all $t \in [u_1, u_2]$ versus various alternative hypotheses. For example, for a general alternative hypothesis, the test statistic $\sup_{u_1 \leq t \leq u_2} |\hat{\eta}(t)/\hat{v}(t)|$ can be used, with critical value determined analytically or by simulations. For the example, the test statistic equals 6.98 and the simulation-based p -value is < 0.001 .

The confidence bands constructed here can also be based on nonparametric hazard estimates that use time-varying kernels and time-varying or data-dependent bandwidths (Müller and Wang, 1994). Furthermore, the techniques presented in this article can be used for constructing simultaneous interval estimates for contrasts of two intensity functions of point processes. Applications include cause-specific hazard functions for competing risks data and transition intensities in Markov chain models under general types of censoring beyond the random censorship model. In addition, under a semiparametric regression model (e.g., the Cox proportional hazards model), confidence bands for the hazard function given a set of covariates can be obtained accordingly.

ACKNOWLEDGEMENTS

The authors thank Professors Yanqing Sun and Naisyin Wang for helpful comments and the associate editor and a referee for input that improved the article. This work was supported by NIH grants AI46703-01 (PBG), CA-56844 (LJW), and CA-75142 (MRK).

RÉSUMÉ

Dans les analyses de survie, l'évolution du risque instantané dans le temps est souvent plus intéressante que celle du risque cumulé. Dans ce cas, un estimateur non paramétrique de la fonction de risque est plus approprié pour résumer l'évolution du risque que l'estimateur de Kaplan Meier correspondant. Lorsqu'on compare un nouveau traitement à un traitement de référence, il est important de savoir si le traitement perd de son efficacité durant le suivi, et si oui à partir de quel moment. Malheureusement avec un graphique montrant la différence de courbes de Kaplan Meier, il est plutôt difficile de visualiser de telles évolutions temporelles. Dans cet article, nous proposons des méthodes simples pour construire des intervalles de confiance autour de la différence de deux fonctions de risque instantané en présence de données censurées. Les estimateurs simultanés des intervalles de confiance sont très utiles pour identifier des valeurs possibles de la différence au cours du temps avec un certain degré de confiance. Ces nouvelles méthodes sont illustrées à l'aide d'un exemple et d'une petite étude de simulation.

REFERENCES

- Aalen, O. (1978). Nonparametric inference for a family of counting processes. *Annals of Statistics* **6**, 701–726.

- Andersen, P. K., Borgan, O., Gill, R. D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer.
- Bickel, P. J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics* **1**, 1071–1095.
- Clemens, J. D., Sack, D. A., Harris, J. R., et al. (1990). Field trial of oral cholera vaccines in Bangladesh: Results from three-year follow-up. *Lancet* **335**, 270–273.
- Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan–Meier curve. *Journal of the American Statistical Association* **83**, 414–425.
- Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *Journal of the American Statistical Association* **88**, 1287–1301.
- Gasser, T. and Müller, H.-G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation, Lecture Notes in Mathematics*, Volume 757, 23–68. Berlin: Springer-Verlag.
- Gilbert, P. B. and Kosorok, M. R. (2001). *Simultaneous inferences in nonparametric functional estimation*. Technical Report, University of Washington, Seattle, Washington.
- Hall, P. (1991). On convergence rates of suprema. *Probability Theory and Related Fields* **89**, 447–455.
- Hall, P. and Owen, A. B. (1993). Empirical likelihood confidence bands in density estimation. *Journal of Computational and Graphical Statistics* **2**, 273–289.
- Hall, P. and Titterton, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *Journal of Multivariate Analysis* **27**, 228–254.
- Halloran, M. E., Haber, M. J., and Longini, I. M. (1992). Interpretation and estimation of vaccine efficacy under heterogeneity. *American Journal of Epidemiology* **136**, 328–343.
- Härdle, W. and Bowman, A. W. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *Journal of the American Statistical Association* **83**, 102–110.
- Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.
- Lin, D. Y., Fleming, T. R., and Wei, L. J. (1994). Confidence bands for survival curves under the proportional hazards model. *Biometrika* **81**, 73–81.
- Müller, H.-G. and Wang, J.-L. (1994). Hazard rate estimation under random censoring with varying kernels and bandwidths. *Biometrics* **50**, 61–76.
- Nelson, W. (1972). Theory and applications of hazard plotting for censored failure time data. *Technometrics* **14**, 945–966.
- Parzen, M. I., Wei, L. J., and Ying, Z. (1997). Simultaneous confidence intervals for the difference of two survival functions. *Scandinavian Journal of Statistics* **24**, 309–314.
- Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Annals of Statistics* **11**, 453–466.
- Rice, J. and Rosenblatt, M. (1976). Estimation of the log survivor function and hazard function. *Sankhyā, Series A* **38**, 60–78.

- Tanner, M. A. and Wong, W.-H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *The Annals of Statistics* **11**, 989–993.
- Van der Vaart, A. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.
- Xia, Y. (1998). Bias-corrected confidence bands in nonparametric regression. *Journal of the Royal Statistical Society, Series B* **60**, 797–811.
- Yandell, B. S. (1983). Nonparametric inference for rates with censored survival data. *The Annals of Statistics* **11**, 1119–1135.

Received April 2001. Revised May 2002.

Accepted May 2002.