Communications for Statistical Applications and Methods
2017, Vol. 24, No. 2, 115–127

https://doi.org/10.5351/CSAM.2017.24.2.115
Print ISSN 2287-7843 / Online ISSN 2383-4757

Bayesian test for the differences of survival functions in multiple groups

Gwangsu Kim[1,a]

[a]Data Science for Knowledge Creation Research Center, Seoul National University, Korea

Abstract

This paper proposes a Bayesian test for the equivalence of survival functions in multiple groups. Proposed Bayesian test use the model of Cox's regression with time-varying coefficients. B-spline expansions are used for the time-varying coefficients, and the proposed test use only the partial likelihood, which provides easier computations. Various simulations of the proposed test and typical tests such as log-rank and Fleming and Harrington tests were conducted. This result shows that the proposed test is consistent as data size increase. Specifically, the power of the proposed test is high despite the existence of crossing hazards. The proposed test is based on a Bayesian approach, which is more flexible when used in multiple tests. The proposed test can therefore perform various tests simultaneously. Real data analysis of Larynx Cancer Data was conducted to assess applicability.

Keywords: Cox's regression, survival functions, log-rank test, Fleming and Harrington test, Bayes factor, time-varying coefficients

1. Introduction

In a survival analysis, it is a traditional and important problem to identify the differences in survival functions $S_k, k = 1, \ldots, K$, where the S_k denotes the survival function of the k^{th} group. The survival function is not scalar and makes finding differences in survival functions difficult. There are many issues related to this problem. The log-rank test (Mantel, 1966) and the Fleming and Harrington test (Harrington and Fleming, 1982) are commonly used for this.

If we define the hazard function for the k^{th} group as

$$\lambda_k(t) = \lim_{\Delta \downarrow 0} \frac{P\left(t + \Delta > T_k \geq t\right) / P\left(T_k \geq t\right)}{\Delta}, \quad 0 < t < \infty,$$

where the lifetime of k^{th} group follows the distribution of T_k, and temporarily assume that $K = 2$, then log-rank test is very powerful when $0 < \lambda_1(t)/\lambda_2(t) = c < \infty$.

However, in an actual data analysis, proportionality is a strong assumption, and the ratio of hazard functions can vary. Especially, with multiple groups, the possibility of a varying hazard ratio is not rare and greatly reduces the power of log-rank test. Weighted log-rank tests can be used for a varying hazard ratio. The weights of these tests are determined by the Kaplan-Meier estimator (Kaplan and Meier, 1958) of the survival function. The power levels of the tests depend on the weights. Differences in survival functions can occur earlier or later, and the weighted log-rank test can provide more

[1] Data Science for Knowledge Creation Research Center, Seoul National University, 1, Gwanak-ro, Gwanak-gu, Seoul 08826, Korea. E-mail: s88012@gmail.com

Published 31 March 2017 / journal homepage: http://csam.or.kr

weights earlier or later. To overcome this problem, Park and Jeong (1995) considered combinations of weighted log-rank tests and Kim *et al.* (2001) considered random permutation tests and conducted various simulations for various setups of differences of survival functions in two groups. In addition, Muggeo and Tagliavia (2010) and Yang and Prentice (2005) also proposed novel tests for the crossing hazard ratio. Various results of simulations were reported in Li *et al.* (2015). To the best of my knowledge, there are few tests or simulation results for multiple groups with the exception of the (weighted) log-rank test.

In this paper, Cox's model and time-varying coefficients are introduced in Section 2. I propose the test for the equivalence of survival functions in multiple groups in Section 3. Various simulations were conducted to validate the consistency of the proposed test (especially when a crossing hazard exists) in Section 4, and the analysis with real data is conducted in Section 5. Concluding remarks are given in Section 6.

2. Preliminaries of Cox' model with time-varying coefficients

We assume that the lifetime follows the law of a non-negative random variable T having survival function S and that probability distribution of the residual lifetime can be obtained via $P(T > s|T > t) = S(s)/S(t)$, $s \geq t \geq 0$. In addition, S is represented by the cumulative hazard function A by

$$S(t) = \prod_{s\in(0,t]} (1 - dA(s)).$$

Here, A is a monotonic increasing function defined of $[0, \infty)$ such that $A(0) = 0$ and $\lim_{t\to\infty} A(t) = \infty$. If A is absolute continuous, then there exist a hazard function, λ, such that

$$A(t) = \int_0^t \lambda(s)ds \quad \text{and} \quad S(t) = \exp\left(-\int_0^t \lambda(s)ds\right). \tag{2.1}$$

The model of (2.1) does not consider the heterogeneity of hazard functions, but we can introduce a covariate z which takes into account the effects of the hazard function. To do this, we consider Cox's model (Cox, 1972) with time-varying coefficients:

$$\lambda(t|z) = \exp\left(z^T\beta(t)\right)\lambda_0(t), \tag{2.2}$$

where λ_0 is the baseline hazard function and $\beta(t) = (\beta_1(t), \ldots, \beta_{K-1}(t))^T$. If z is an indicator variable such that

$$\underbrace{(0, \ldots, 0)^T}_{K-1} : \text{control group (first group)}, \qquad \underbrace{(0, \ldots, 0}_{k-1}, 1, \underbrace{0, \ldots, 0)^T}_{K-k-1} : (k+1)^{th} \text{ group } (k \leq K-1),$$

then

$$\begin{aligned}
\lambda_0(t) &= \lambda\left(t|z = (0, 0, \ldots, 0)^T\right) = \lambda_0(t),\\
\lambda_1(t) &= \lambda\left(t|z = (1, 0, \ldots, 0)^T\right) = \exp(\beta_1(t))\,\lambda_0(t),\\
&\vdots \qquad\qquad\qquad \vdots \qquad\qquad\qquad \vdots\\
\lambda_{K-1}(t) &= \lambda\left(t|z = (0, 0, \ldots, 1)^T\right) = \exp(\beta_{K-1}(t))\,\lambda_0(t),
\end{aligned}$$

where λ_{k-1} is the hazard function of the k^{th} group. Thus, testing of $\beta_1(\cdot) = \cdots = \beta_{K-1}(\cdot) \equiv 0$ means the testing of the equivalence of hazard and survival functions between all groups.

Meanwhile, the lifetime is usually right-censored by a censoring variable C, where T and C are conditionally independent given z. Observed data are from a random sample, data $\stackrel{\triangle}{=} \{X_i, z_i, \delta_i\}_{i=1}^n$, where

$$\begin{aligned} X_i &= \min(T_i, C_i), \qquad \delta_i = I(T_i \le C_i), \\ P(T_i > t|z_i) &= \exp\left(-\int_0^t \lambda(s|z_i)ds\right), \\ C_i|z_i &\sim G_i, \quad C_i \perp T_i \text{ conditionally on } z_i, \end{aligned}$$

and G_i is the distribution function of C_i.

3. Bayesian test for the equivalence of survival functions in multiple groups

3.1. Bayesian test

First, we review the Bayesian test. We assume that $D_{1:n} \stackrel{\triangle}{=} \{X_i\}_{i=1}^n$ is generated from $f(\cdot|\theta_{k^*}, M_{k^*})$ where $k^* \in \{1, \ldots, K\}$. In addition, we let π_k and p_k be a prior for θ_k and a prior mass for model M_k, i.e., $\pi(M_k)$. Then posterior of each M_k is

$$\pi(M_k|D_{1:n}) = \frac{p_k \int \left[\prod_{i=1}^n f(X_i|\theta_k, M_k)\right] \pi_k(d\theta_k|M_k)}{\sum_{s=1}^K p_s \int \left[\prod_{i=1}^n f(X_i|\theta_s, M_s)\right] \pi_s(d\theta_s|M_s)}. \tag{3.1}$$

The Bayesian test is conducted using these values. If $\pi(M_{k^*}|D_{1:n}) \to 1$, and $\pi(M_k|D_{1:n}) \to 0$ for $k \neq k^*$, it implies the consistency of model selection. If this consistency is ensured, we can use these posteriors of (3.1) in tests. Related to this approach, Bayes factors are used for model selection (Kass and Raftery, 1995). Bayes factors are as:

$$B_{k_2k_1} = \frac{\int \left[\prod_{i=1}^n f(X_i|\theta_{k_2}, M_{k_2})\right] \pi_{k_2}(d\theta_{k_2}|M_{k_2})}{\int \left[\prod_{i=1}^n f(X_i|\theta_{k_1}, M_{k_1})\right] \pi_{k_1}(d\theta_{k_1}|M_{k_1})},$$

where $(k_1, k_2) \in \{1, \ldots, K\}^2$. Note that

$$\frac{\pi(M_{k_2}|D_{1:n})}{\pi(M_{k_1}|D_{1:n})} = \frac{B_{k_2k_1}\pi(M_{k_2})}{\pi(M_{k_1})}, \tag{3.2}$$

and $B_{k^*k} \to \infty$ $(k \neq k^*)$ is equal to $\pi(M_{k^*}|D_{1:n}) \to 1$ when $0 < \min_k \pi(M_k) \le \max_k \pi(M_k) < 1$.

3.2. Partial likelihood and Bayesian test

In the survival analysis, partial likelihood is often used instead of the probability density function. The asymptotic properties of partial likelihood were studied by Andersen and Gill (1982), Tsiatis (1981) and others. If we consider the time-varying coefficients model, the partial likelihood of Cox's model with time-varying coefficients is defined as

$$L^p(\beta; \text{data}) = \prod_{i=1}^n \left[\frac{\exp\left(z_i^T\beta(X_i)\right)}{\sum_{j:X_j \ge X_i} \exp\left(z_j^T\beta(X_i)\right)}\right]^{\delta_i}$$

for data. It is assumed that there exists a $0 < \tau < \infty$ such that $G_i(\tau) = 1$, and $G_i(\tau-) < 1$. Thus, $P(X_i \le \tau) = 1$.

In the model of (2.2), if the β_ks are constant function, the proportional hazards assumption holds, whereas if β_k functions are monotonic functions crossing 1, then crossing hazards exist. For β_k, we use the B-spline basis expansion such that $\sum_{j=1}^{a_n} \gamma_{k,j} B_{a_n,j}$, where the $B_{a_n,j}$s are the B-spline basis functions with equally spaced knots, and a_n is a diverging sequence as $n \to \infty$.

Note that the β_ks in a certain function class such that (3.3) are approximated by B-spline basis functions (sieve approach). For the test, we consider the model of $\eta_k \sum_{l=1}^{a_n} \gamma_{k,l} B_{a_n,l}$, and priors such that

$$\begin{aligned} \pi\left(d\gamma_{1,1} \cdots d\gamma_{K-1,a_n}\right) = & \left[\prod_{k=1}^{K-1} \prod_{l=1}^{a_n} h(\gamma_{k,l})\right] d\gamma_{1,1} \cdots d\gamma_{K-1,a_n}, \\ \eta_k | p_k \sim & \text{ Bernoulli}(p_k), \\ p_k \sim & \text{ Beta}(1, \alpha_k), \quad \alpha_k > 0. \end{aligned}$$

Here, $\forall (k,l) \in \{1, \ldots, K-1\} \times \{1, \ldots, a_n\}$, $\pi\left(|\gamma_{k,l}| > B\right) = 0$, and

$$0 < c_* \le \inf_{\gamma \in [-B,B]} h(\gamma) \le \sup_{\gamma \in [-B,B]} h(\gamma) \le c^* < \infty.$$

The value of $B > 0$ is sufficiently large compared to $L > 0$ in (3.3), and α_k is a hyper-parameter resulting in $\pi(\eta_k = 1) = 1/(1+\alpha_k)$. Related to this approach, Kim and Lee (2016) proposed a Bayesian test using the partial likelihood and time-varying coefficients model. They let $K = 2$, $a_n = O([(n/\log n)^{1/(2p+1)}])$, and define

$$\begin{aligned} \Theta_p^L &= \left\{ \beta : \sup_{t \in [0,\tau]} \left|\beta^{(m)}(t)\right| < L, \ m = 0, \ldots, p \right\}, \\ \mathcal{F}_0 &= \{\beta : \beta(\cdot) \equiv 0\}, \end{aligned} \tag{3.3}$$

where $\beta^{(m)}$ is the m^{th} derivative function of β and $\beta^{(0)} = \beta$. With this setup, Kim and Lee (2016) claimed the consistency of the test

$$H_0 : \beta_1 \in \mathcal{F}_0 \quad \text{vs.} \quad H_1 : \beta_1 \in \Theta_p^L \setminus \mathcal{F}_0,$$

i.e., $\pi(\eta_1 = 0 |$ data $) \to 1$ when data are generated from β_1 in H_0 and $\pi(\eta_1 = 1 |$ data $) \to 1$ when data are generated from β_1 in H_1. Thorough theoretical studies of this test are found in Kim *et al.* (2017).

Now we let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{K-1})^T$ and consider various test such that

$$\begin{aligned} H_0 : \lambda_1 = \cdots = \lambda_{K-1} = \lambda_0 \quad \text{vs.} \quad H_1 : \text{ not } H_0, \\ H_0 : \lambda_k = \lambda_0 \quad \text{vs.} \quad H_1 : \text{ not } H_0, \end{aligned}$$

and $H_0 : \lambda_1 = \lambda_{K-1} = \lambda_0$ vs. not H_0. Then, $\pi(\boldsymbol{\eta} \neq \mathbf{0} |$ data $)$, $\pi(\eta_k \neq 0 |$ data $)$ and $\pi((\eta_1, \eta_{K-1}) \neq (0,0) |$ data $)$ can be used for these tests. We consider the typical test of

$$H_0 : \lambda_1 = \cdots = \lambda_{K-1} = \lambda_0 \quad \text{vs.} \quad H_1 : \text{ not } H_0.$$

In this test $\pi(\boldsymbol{\eta} \neq \mathbf{0} |$ data $)$ can be obtained from data, and it constitute a test statistic. If we have a consistent Bayesian test, then we need criteria with which to reject H_0 in a finite sample. We consider

Table 1: Setups for simulated data

	M_0	M_{1c}	M_{1m}
First group	$\lambda_0(t) = 0.25$	$\lambda_0(t) = 0.25$	$\lambda_0(t) = 0.25$
Second group	$\lambda_1(t) = 0.25$	$\lambda_1(t) = 0.25$	$\lambda_1(t) = 0.25$
Third group	$\lambda_2(t) = 0.25$	$\lambda_2(t) = 0.25\exp(0.1 + 1.5t)$	$\lambda_2(t) = 0.25\exp(3.0 + 1.5t)$

three decision rules. The first is that $\pi(\boldsymbol{\eta} \neq \mathbf{0}|\text{ data }) > 0.5$ denoted by C_m, and the second is that

$$\frac{\pi(\boldsymbol{\eta} \neq \mathbf{0}|\text{ data })}{\pi(\boldsymbol{\eta} = \mathbf{0}|\text{ data })} > \frac{\pi(\boldsymbol{\eta} \neq \mathbf{0})}{\pi(\boldsymbol{\eta} = \mathbf{0})},$$

denoted by C_p. The third rule is that of Kass and Raftery (1995). They held that if $\log_{10}(B_{H_1H_0})$ exceeds 1/2, then H_1 is substantial. If we let $K = 3$, this criterion with (3.2) implies that the value of

$$\frac{\pi(\boldsymbol{\eta} \neq \mathbf{0}|\text{ data })}{\pi(\boldsymbol{\eta} = \mathbf{0}|\text{ data })}$$

exceeds 9.49 when $\pi(H_0) = \pi(\boldsymbol{\eta} = \mathbf{0}) = 0.25$. This implies that $\pi(\boldsymbol{\eta} \neq \mathbf{0}|\text{ data }) > 9.49/10.49 \doteq 0.90$. Moreover, in the test of $H_0 : \lambda_k = \lambda_0$ vs. $H_1 :$ not H_0, the rejection criterion of H_0 is that $\pi(\eta_k = 1|\text{ data })$ exceeds 0.76 when $\pi(\eta_k = 1) = 0.50$. We use the notation of C_B for Kass and Raftery's (1995) criterion. We observed the behaviors of all approaches in simulation studies.

The advantage of the Bayesian test is that various tests can be done simultaneously, making it superior to other tests which perform only one test, such as $H_0 : \lambda_1 = \cdots = \lambda_{K-1} = \lambda_0$ vs. $H_1 :$ not H_0 and $H_0 : \lambda_k = \lambda_0$ vs. $H_1 :$ not H_0. From the proposed test, we can simultaneously identify groups that have differences and those that do not.

3.2.1. Computations

We use truncated normal distributions for the hs, and a joint distribution (using partial likelihood) of the random sample and random parameters along with the posterior of the random parameters are in the Appendix A. In addition, Hastings (1970) and Metropolis *et al.* (1953) are referred to for the algorithm with which to obtain posteriors. This paper used rejection sampling; however, more advance studies can be found in Gilks and Wild (1992) and in Kim and Lee (2003).

4. Simulation results

4.1. Setups

In the simulations, three groups are considered. All three hazards are equal in the M_0 setup. In the M_{1c} setup, one hazard ratio with respect to control group is 1 and the hazard ratio is crossing. In the M_{1m} setup, one hazard ratio is 1 and other hazard ratio is monotonically surpassing over 1. In the simulations, all censoring variables are generated from exponential distributions truncated by 6, which creates censoring rates of 0.25, 0.50, and 0.75. The data sizes considered are 90, 150, and 300, where the data sizes of all groups are equal. The hazard functions of the simulated data (with 100 replications) are summarized in Table 1.

In B-spline basis expansion, we use five basis functions with a degree of 2, and use priors of truncated normals for $\gamma_{k,l}$ and let $\alpha_1 = \alpha_2 = 1$. Note that $\pi(\eta_1 = \eta_2 = 0) = \pi(\eta_1 = 0)\pi(\eta_2 = 0) = 0.5^2 = 0.25$, and the joint distribution and posterior are provided in Appendix A. From MCMC (Markov chain Monte Carlo) we obtained 6,000 chains. We discarded the first 1,000 chains. We

Table 2: Ratio of rejection of $\lambda_1 = \lambda_2 = \lambda_0$ in the proposed test and others when data size are 90, 150, and 300, respectively

Censoring rates	Methods	Data size 90			Data size 150			Data size 300		
		M_0	M_{1c}	M_{1m}	M_0	M_{1c}	M_{1m}	M_0	M_{1c}	M_{1m}
0.25	Proposed (C_B)	0.04	0.63	0.89	0.03	0.97	1.00	0.04	1.00	1.00
	Proposed (C_p)	0.12	0.82	0.95	0.07	0.99	1.00	0.09	1.00	1.00
	Proposed (C_m)	0.30	0.97	1.00	0.20	1.00	1.00	0.14	1.00	1.00
	Log-rank	0.06	0.57	0.95	0.06	0.83	1.00	0.06	1.00	1.00
	Fleming & H(1)	0.03	0.24	0.81	0.08	0.36	1.00	0.04	0.59	1.00
	Fleming & H(−1)	0.07	0.92	0.99	0.06	1.00	1.00	0.09	1.00	1.00
0.50	Proposed (C_B)	0.05	0.34	0.58	0.03	0.65	0.90	0.06	0.99	0.99
	Proposed (C_p)	0.17	0.56	0.80	0.13	0.87	0.95	0.09	0.99	1.00
	Proposed (C_m)	0.38	0.71	0.92	0.24	0.95	0.99	0.14	1.00	1.00
	Log-rank	0.06	0.26	0.75	0.05	0.46	0.96	0.03	0.80	1.00
	Fleming & H(1)	0.05	0.11	0.60	0.05	0.18	0.87	0.03	0.37	0.98
	Fleming & H(−1)	0.08	0.51	0.83	0.09	0.85	0.96	0.08	0.97	1.00
0.75	Proposed (C_B)	0.02	0.12	0.28	0.08	0.28	0.47	0.01	0.57	0.82
	Proposed (C_p)	0.20	0.40	0.52	0.17	0.47	0.70	0.10	0.73	0.89
	Proposed (C_m)	0.42	0.67	0.80	0.34	0.76	0.85	0.25	0.87	0.96
	Log-rank	0.05	0.08	0.39	0.08	0.14	0.65	0.02	0.29	0.90
	Fleming & H(1)	0.04	0.06	0.32	0.09	0.06	0.59	0.04	0.19	0.85
	Fleming & H(−1)	0.08	0.15	0.43	0.10	0.22	0.70	0.03	0.53	0.91

Fleming and Harrington test with $\rho = 1$ and Fleming and Harrington test with $\rho = -1$ are abbreviated to Fleming & H(1) and Fleming & H(−1), respectively.

obtained a posterior of size 200 by the 25^{th} thinning of the remaining chains. Figure B.1 in the Appendix B shows the cumulative means of the posterior in a specific setup, which implies that the ergodic condition is satisfied.

We compared the proposed test with the popular log-rank test and with Fleming and Harrington test extended to multiple groups (the weights are $\hat{S}(t)^{\rho},\ \rho \in \{1, -1\}$). For these tests, we use the function of SURVDIFF in R, and significance levels are set to 0.05 for the rejection of H_0. The frequentist test and the Bayesian test are different; in addition, there is no common baseline with which to compare these tests. Comparisons with other tests are only for the behavior of the proposed test, and not to claim better performance.

4.2. Results

4.2.1. Comparison with other tests

Table 2 summarizes the rejection ratios observed in the simulations. It is obvious that the Fleming and Harrington tests are very sensitive to the weights. In the M_{1c} setup, the Fleming and Harrington test of $\rho = 1$ has lower power while that of $\rho = -1$ has higher power. This can be interpreted as indicating that an earlier crossing later requires more weight for higher power. Note that the differences of survival functions increase with time in the M_{1c} setup, it means that Fleming and Harrington test of $\rho = 1$ gives lower weights later, which results in the lower power levels. Meanwhile, the power of the log-rank test is between the Fleming and Harrington test with $\rho = -1$ and $\rho = 1$. In the M_{1m} setup, power levels of the log-rank test and the Fleming and Harrington tests are higher than those of the M_{1c} setup, but the trends are similar to the M_{1c} setup.

Obviously the power levels decrease with higher censoring. The Fleming and Harrington test of $\rho = -1$ has a somewhat higher value than the nominal coverage rate of 0.05 in the M_0 setup. This can be interpreted as indicative of a trade-off in the test.

The power levels in tests increase in all setups as the data size increases. When the data size is

300, power levels of the log-rank tests and the Fleming and Harrington tests decrease in the M_{1c} setup as the censoring rates increase, while they become more steadily lower in the M_{1m} setup. This implies that traditional tests have lower power when there are crossing hazards and high censoring. However, the power levels of these tests are high in the M_{1m} setup when the sample size increases despite high censoring.

In the proposed test (C_B), the correct selection ratio of M_0 are slightly larger than in other tests. When the data size is 90, the power of the proposed test (C_B) surpasses the log-rank test and the Fleming and Harrington test with $\rho = 1$ in M_{1c} and exceeds the Fleming and Harrington test with $\rho = 1$ at the censoring rates of 0.25 and M_{1m}. The power levels of the proposed test (C_B) become higher as the data size increase. The power levels of the proposed test (C_B) are higher or nearly equal to those in other tests with the exception of a censoring rate of 0.75 in M_{1m} when the data size is 150. The behaviors are very similar when the data size is 300. Compared to the case of a data size of 150, the differences in the power levels subside slightly in the M_{1m} setup. Note that C_p and C_m are not recommended because the false selection ratios are too large in the M_0 setup.

The proposed test is based on a nonparametric approach, which results in less gradual convergence given a small sample size. Furthermore, high censoring degrades the efficiency of B-spline approximation because differences in the functions can be found in non-censoring time points when using partial likelihood. These outcomes explain the lower power in the event of a small sample size and high censoring.

The proposed test result in the less lower with small sample size and the M_{1m} setup, but the proposed test outperformed nearly all other tests in the M_{1c} setup, only the Fleming and Harrington test with $\rho = -1$ can be comparable. In some cases (sample size 90, sample size 150 and censoring rate <= 0.50), the Fleming and Harrington test with $\rho = -1$ performs better, but the proposed test performs very well with an increase in the sample size.

4.2.2. Results of marginal tests

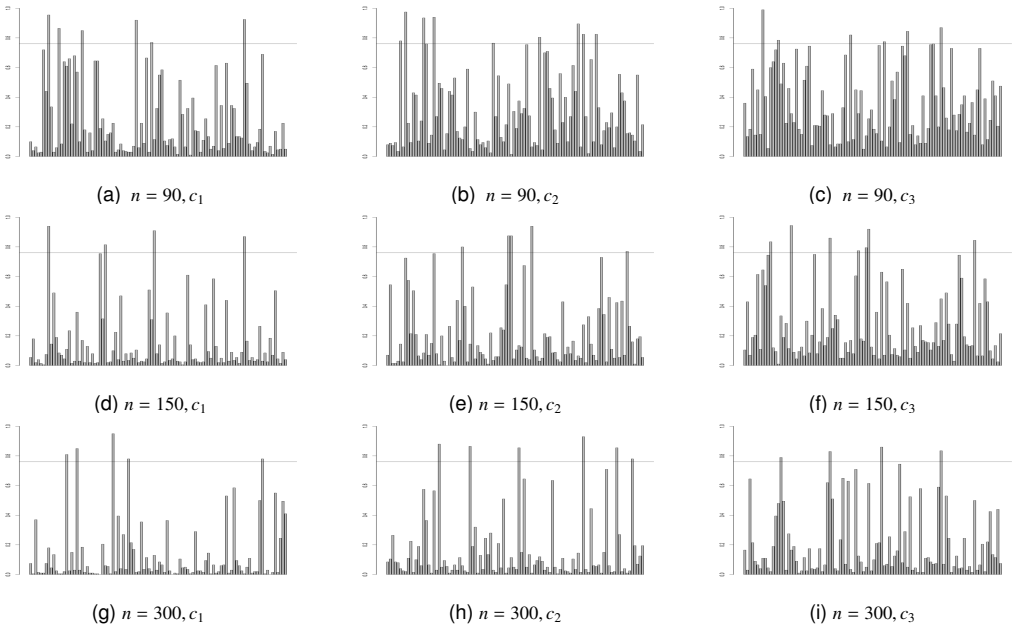
According to the proposed test, we can test the difference between a certain group with a control group. If we consider the two tests of

$$\begin{aligned} T_1; \quad & H_0 : \lambda_1 = \lambda_0 \quad \text{vs.} \quad H_1 : \text{not } H_0, \\ T_2; \quad & H_0 : \lambda_2 = \lambda_0 \quad \text{vs.} \quad H_1 : \text{not } H_0, \end{aligned}$$

then the proposed test can perform the two tests using posterior samples, the results of the M_0 setup are summarized in Figures 1 and 2. These plots show that the ratios of false selections (using C_B) are around 0.02–0.05.

In addition, Table 3 summarizes the results of the M_{1c} and M_{1m} setups. It is obvious that $\pi(\eta_1 = 0|\text{ data})$ and $\pi(\eta_2 = 1|\text{ data})$ increases with an increase in sample sizes. The acceptance ratio of $\lambda_1 = \lambda_0$ are around 0.95 and 0.85, and rejection ratio of $\lambda_2 = \lambda_0$ depends on the sample size and censoring rate. It shows a value of 1.00–0.50 in the most setups. The rejection ratio of $\lambda_2 = \lambda_0$ increases as the sample size increases and the censoring rate decreases. If the sample size exceeds 90 and the censoring rate is less than or equal to 0.50, then the rejection ratio of $\lambda_2 = \lambda_0$ is greater than or equal to 0.82. In addition, the rejection ratio of $\lambda_2 = \lambda_0$ in the M_{1m} setup is larger than in the M_{1c} setup. Finally, we ensure that the value of $\pi(\eta_2 = 1|\text{ data})$ goes to 1 if the hazard ratio is far from 1, the sample size is large, and the censoring rate is relatively lower. This demonstrates the consistency of the proposed test.

The posterior of $\eta_1 = 0$ is not affected by the censoring rates due to the equivalence of hazards. Thus, the power levels of the proposed test are affected by the hazard ratio.

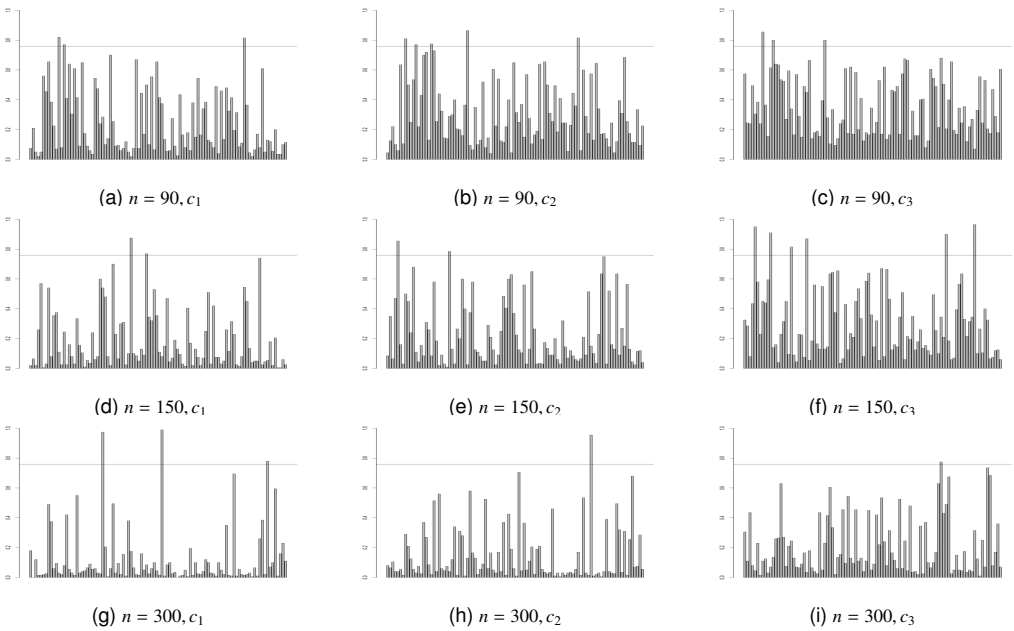


(a) $n = 90, c_1$ (b) $n = 90, c_2$ (c) $n = 90, c_3$

(d) $n = 150, c_1$ (e) $n = 150, c_2$ (f) $n = 150, c_3$

(g) $n = 300, c_1$ (h) $n = 300, c_2$ (i) $n = 300, c_3$

Figure 1: *Plots of $\pi(\eta_1 = 1|$ data $)$ of 100 replications in the M_0 setup. The c_1, c_2, and c_3 represent the censoring rates of 0.25, 0.50, and 0.75, respectively, and all values of bars are 0.76.*



(a) $n = 90, c_1$ (b) $n = 90, c_2$ (c) $n = 90, c_3$

(d) $n = 150, c_1$ (e) $n = 150, c_2$ (f) $n = 150, c_3$

(g) $n = 300, c_1$ (h) $n = 300, c_2$ (i) $n = 300, c_3$

Figure 2: *Plots of $\pi(\eta_2 = 1|$ data $)$ of 100 replications in the M_0 setup. The c_1, c_2, and c_3 represent the censoring rates of 0.25, 0.50, and 0.75, respectively, and all values of bars are 0.76.*

Table 3: Ratio of $\pi(\eta_1 = 1|\text{ data }) > 0.76$ (denoted by r_1) and $\pi(\eta_2 = 1|\text{ data }) > 0.76$ (denoted by r_2) for all setups of M_{1c} and M_{1m}

Data sizes	Censoring rates	r_1		r_2	
		M_{1c}	M_{1m}	M_{1c}	M_{1m}
90	0.25	0.08	0.08	0.79	0.93
	0.50	0.13	0.15	0.49	0.65
	0.75	0.13	0.13	0.15	0.39
150	0.25	0.12	0.11	0.99	1.00
	0.50	0.12	0.05	0.82	0.95
	0.75	0.13	0.14	0.33	0.68
300	0.25	0.13	0.08	1.00	1.00
	0.50	0.13	0.08	0.99	1.00
	0.75	0.08	0.06	0.68	0.89

4.3. Sensitivity analysis

In this section, sensitivity analyses were done with respect to a_n and α_k. We used five B-spline basis functions of order 2, requiring four inner knots. It is near the minimal number to capture the non-linear structure in practical data analysis. In the sensitivity analysis, five basis functions / seven basis functions and $\alpha_k = 1$ / $\alpha_k = 2$ are considered. In addition, we consider the censoring rates of 0.25 and 0.50, adequate to observe the overall trends. We observed rejection ratio of $H_0 : \lambda_0 = \lambda_1 = \lambda_2$. Table B.1 in Appendix B summarizes the results. The number of B-spline basis functions does not greatly effect the performance of the proposed test, but the effects of α_k are strong when data the size is small and the censoring rate is high. These results appear to be natural because α_k is directly related to the model priors. It is possible that a data size of 150 and a censoring rate 0.25–0.50 may mitigate the affect of the prior.

5. Real data analysis

In the real data analysis, we consider Larynx Cancer Data from 90 patients. Among the 90 patients, incidences of stages I, II, III, and IV cancer are 33, 17, 27, and 13 at the beginning of therapy. The censoring rate is 0.44. Figure 3 represents the Kaplan-Meier estimate for each stage, showing that the difference between stage I and stage II is not large, and that the survival probabilities abruptly decrease at stages of III and IV. This figure also implies that the changes of the hazard ratios may not be large.

A log-rank test of the equivalence of survival functions rejects the null hypothesis with a p-value of less than 0.0001. In the proposed test, the number and degree of B-spline basis functions and α_k are equal to those in the simulations. Then, we have

$$\pi(\eta_i = 1) = 0.5, \quad i = 1, 2, 3, \quad \text{and} \quad \pi(\eta_1 = a, \eta_2 = b, \eta_3 = c) = \pi(\eta_1 = a)\pi(\eta_2 = b)\pi(\eta_3 = c),$$

where $(a, b, c) \in \{0, 1\}^3$. The η_1, η_2, and η_3 are independent and $\pi(\sum_{i=1}^3 \eta_i = 0) = 0.125$. Here, stage I is the control group and the survival function of the $(k + 1)^{th}$ stage is equal to that of control group is modeled by $\eta_k = 0$. We obtain 455,000 chains. The first 80,000 chains are discarded, and obtain a posterior of size 5,000 by the 75^{th} thinning of the remaining chains. Figure B.2 in the Appendix B shows the cumulative means of the posteriors of η_1, η_2, and η_3. All posteriors appear stationary. Based on the criterion of C_B, we can reject the null hypothesis that all hazards are equal when $\pi(\sum_{k=1}^3 \eta_k > 0|\text{ data }) > 0.96$. In the marginal test for stage II, III, and IV, we reject the null hypothesis H_0 when $\pi(\eta_k = 1|\text{ data }) > 0.76$.

The results of the proposed test indicate that $\pi(\sum_{i=1}^3 \eta_i > 0|\text{ data }) = 0.9975$, $\pi(\eta_1 = 1|\text{ data }) =$

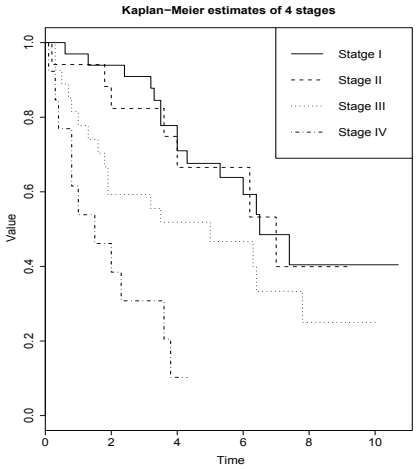


Figure 3: *Plots of Kaplan-Meier estimates of 4 stages of Larynx Cancer Data.*

0.2945, $\pi(\eta_2 = 1|\text{ data }) = 0.7582,$ and $\pi(\eta_3 = 1|\text{ data }) = 0.9943$. Thus, we reject the null hypothesis that all hazards are equal. In marginal tests, we only reject the null hypothesis with the hazards of stage IV equal to that of stage I. We also conducted log-rank tests with stage I and stage II, stage I and stage III, and stage I and stage IV. The p-values from these tests are 0.8663, 0.0800, and less than 0.0001, respectively, showing that the results of the log-rank test and the proposed test are equal. However, it appears that the power of the proposed test is higher than the pairwise log-rank tests because $\pi(\eta_2 = 1|\text{data}) = 0.7582$ is very close 0.76. This may have occurred because the proposed test used all the data whereas the pairwise log-rank tests used partial data in each pair.

6. Concluding remarks

This paper proposed a Bayesian test for differences in hazard and survival functions. Numerical studies and an actual data analysis show the consistency of the proposed Bayesian test and some of its properties. The proposed Bayesian test worked well compared to typical tests; however, the alternative contains crossing hazards. The power of the proposed Bayesian test is nearly equal to typical tests or somewhat lower with a small sample size and higher censoring when the alternative is similar to the degree of the proportionality. Nonetheless, the proposed Bayesian test has the advantages of simultaneous tests and can therefore be very useful in the analysis of actual data.

Meanwhile, the effects of priors cannot be small with a small sample size. More advanced studies of priors of the number of B-spline basis functions and the locations of knots are therefore necessary. These can be done as future works.

Appendix A: Joint distribution (using the partial likelihood) and posteriors

1. Joint distribution (using the partial likelihood) is proportional to

$$L^p\left(\left(\eta_1 \sum_{l=1}^{a_n} \gamma_{1,l} B_{a_n,l}, \eta_2 \sum_{l=1}^{a_n} \gamma_{2,l} B_{a_n,l}, \ldots, \eta_{K-1} \sum_{l=1}^{a_n} \gamma_{K-1,l} B_{a_n,l}\right); \text{data}\right)$$
$$\times \left[\prod_{k=1}^{K-1} \prod_{l=1}^{a_n} N\left(\gamma_{k,l}|0,\sigma^2\right) I\left(|\gamma_{k,l}| < B\right)\right]\left[\prod_{k=1}^{K-1} \text{Ber}\left(\eta_k|p_k\right) \text{Be}\left(p_k|1,\alpha_k\right)\right],$$

where $N(\cdot|\mu, \sigma^2)$, $\text{Ber}(\cdot|p)$ and $\text{Be}(\cdot|a, b)$ denote the probability density functions of the normal distribution with mean μ and variance σ^2, Bernoulli distribution with parameter p, and beta distribution with parameters a and b (mean of $a/(a+b)$), respectively.

2. Posteriors

   (a) For $\gamma_{k,l}|$others,

$$\pi(\gamma_{k,l}|\text{others}) \propto L^p\left(\left(\eta_1 \sum_{l=1}^{a_n} \gamma_{1,l} B_{a_n,l}, \eta_2 \sum_{l=1}^{a_n} \gamma_{2,l} B_{a_n,l}, \ldots, \eta_{K-1} \sum_{l=1}^{a_n} \gamma_{K-1,l} B_{a_n,l}\right); \text{data}\right) \times N\left(\gamma_{k,l}|0, \sigma^2\right) I\left(|\gamma_{k,l}| < B\right), \quad (k,l) \in \{1, \ldots, K-1\} \times \{1, \ldots, a_n\}.$$

   (b) For $\eta_k|$others, $\pi(\eta_k = 1|\text{others}) = A_n/(A_n + B_n)$ where

$$A_n = p_k L^p\left(\left(\eta_1 \sum_{l=1}^{a_n} \gamma_{1,l} B_{a_n,l}, \eta_2 \sum_{l=1}^{a_n} \gamma_{2,l} B_{a_n,l}, \ldots, \eta_{K-1} \sum_{l=1}^{a_n} \gamma_{K-1,l} B_{a_n,l}\right); \text{data}\right)\Bigg|_{\eta_k=1},$$

$$B_n = (1-p_k) L^p\left(\left(\eta_1 \sum_{l=1}^{a_n} \gamma_{1,l} B_{a_n,l}, \eta_2 \sum_{l=1}^{a_n} \gamma_{2,l} B_{a_n,l}, \ldots, \eta_{K-1} \sum_{l=1}^{a_n} \gamma_{K-1,l} B_{a_n,l}\right); \text{data}\right)\Bigg|_{\eta_k=0}.$$

   (c) For $p_k|$others, $\pi(p_k|\text{others}) = Be\left(p_k|1 + \eta_k, \alpha_k + (1 - \eta_k)\right)$.

Appendix B: Plots of posteriors and sensitivity analysis

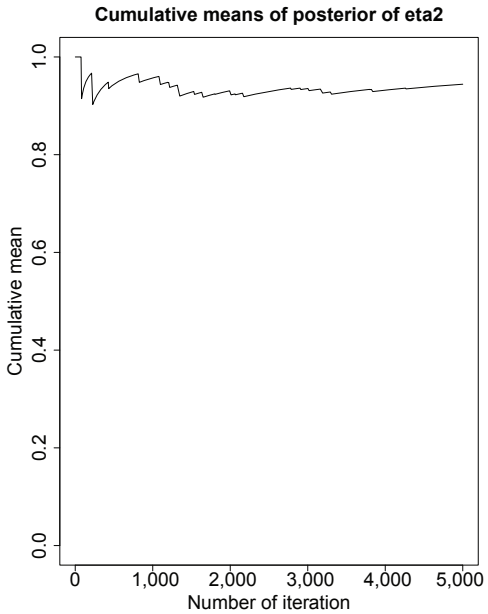


Figure B.1: *Cumulative means of the posterior of η_2 in one replication. These are obtained before thinning with the M_{1c} setup, data size of 150 and 0.25 censoring rate.*

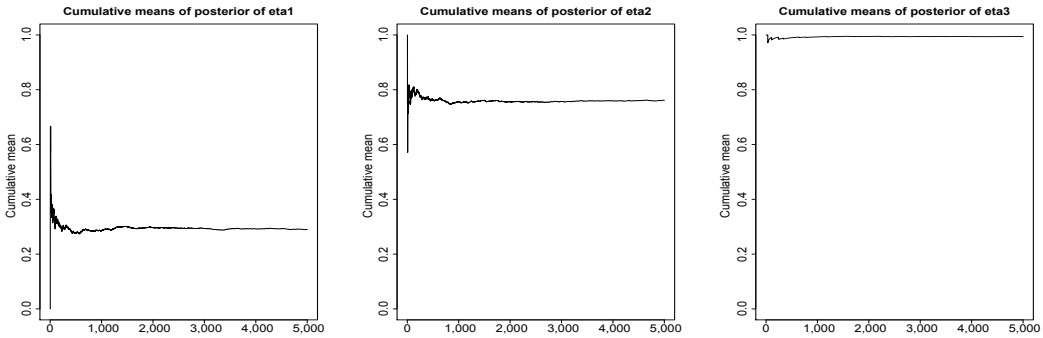


Figure B.2: *Cumulative means of the posteriors of η_1, η_2, and η_3 in Larynx Cancer Data.*

Table B.1: Sensitivity analysis of the proposed test (C_B)

Censoring rates	Data sizes	Setups	M_0	M_{1c}	M_{1m}	Setups	M_0	M_{1c}	M_{1m}
0.25	90	$\alpha = 1$ $a_n = 5$	0.04	0.63	0.89	$\alpha = 1$ $a_n = 7$	0.04	0.67	0.88
	150		0.03	0.97	1.00		0.02	0.96	1.00
	300		0.04	1.00	1.00		0.02	1.00	1.00
0.50	90	$\alpha = 1$ $a_n = 5$	0.05	0.34	0.58	$\alpha = 1$ $a_n = 7$	0.06	0.36	0.57
	150		0.03	0.65	0.90		0.02	0.70	0.89
	300		0.06	0.99	0.99		0.04	0.99	0.98
0.25	90	$\alpha = 2$ $a_n = 5$	0.00	0.21	0.62	$\alpha = 2$ $a_n = 7$	0.00	0.29	0.61
	150		0.00	0.82	0.96		0.00	0.84	0.96
	300		0.00	1.00	1.00		0.01	1.00	1.00
0.50	90	$\alpha = 2$ $a_n = 5$	0.00	0.02	0.26	$\alpha = 2$ $a_n = 7$	0.06	0.06	0.24
	150		0.00	0.27	0.63		0.00	0.37	0.64
	300		0.00	0.90	0.96		0.00	0.91	0.95

The values are the rejection ratios of $H_0 : \lambda_0 = \lambda_1 = \lambda_2$, where $\alpha = \alpha_1 = \alpha_2$.

References

Andersen PK and Gill RD (1982). Cox's regression model for counting processes: a large sample study, *The Annals of Statistics*, **10**, 1100–1120.

Cox DR (1972). Regression models and life-tables, *Journal of the Royal Statistical Society Series B (Methodological)*, **34**, 187–220.

Gilks WR and Wild P (1992). Adaptive rejection sampling for Gibbs sampling, *Applied Statistics*, **41**, 337–348.

Harrington DP and Fleming TR (1982). A class of rank test procedures for censored survival data, *Biometrika*, **69**, 553–566.

Hastings WK (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrika*, **57**, 97–109.

Kaplan EL and Meier P (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association*, **53**, 457–481.

Kass RE and Raftery AE (1995). Bayes factors, *Journal of the American Statistical Association*, **90**, 773–795.

Kim G, Kim Y, and Choi T (2017). Bayesian analysis of the proportional hazards model with time-varying coefficients, *Scandinavian Journal of Statistics*, Advance online publication, doi: 10.1111/sjos.12263

Kim G and Lee S (2016). Bayesian test for hazard ratio in survival analysis, *SpringerPlus*, **5**, 649.

Kim MK, Lee JW, and Huh MH (2001). Random permutation test for comparison of two survival curves, *Communications for Statistical Applications and Methods*, **8**, 137–145.

Kim Y and Lee J (2003). Bayesian bootstrap for proportional hazards models, *The Annals of Statistics*, **31**, 1905–1922.

Li H, Han D, Hou Y, Chen H, and Chen Z (2015). Statistical inference methods for two crossing survival curves: a comparison of methods, *PLoS One*, **10**, e0116774.

Mantel N (1966). Evaluation of survival data and two new rank order statistics arising in its consideration, *Cancer Chemotherapy Reports*, **50**, 163–170.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, and Teller E (1953). Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, **21**, 1087–1092.

Muggeo VMR and Tagliavia M (2010). A flexible approach to the crossing hazards problem, *Statistics in Medicine*, **29**, 1947–1957.

Park SG and Jeong GJ (1995). On combination of several weighted logrank tests, *Communications for Statistical Applications and Methods*, **2**, 213–220.

Tsiatis AA (1981). A large sample study of Cox's regression model, *The Annals of Statistics*, **9**, 93–108.

Yang S and Prentice R (2005). Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data, *Biometrika*, **92**, 1–17.

Received September 26, 2016; Revised January 27, 2017; Accepted February 4, 2017