Statistics in Biopharmaceutical Research

ISSN: (Print) 1946-6315 (Online) Journal homepage: https://amstat.tandfonline.com/loi/usbr20

Hazard Ratio Estimation in Small Samples

Rengyi Xu, Pamela A. Shaw & Devan V. Mehrotra

To cite this article: Rengyi Xu, Pamela A. Shaw & Devan V. Mehrotra (2018) Hazard Ratio Estimation in Small Samples, Statistics in Biopharmaceutical Research, 10:2, 139-149, DOI: 10.1080/19466315.2017.1369899

To link to this article: https://doi.org/10.1080/19466315.2017.1369899

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2018, VOL. 10, NO. 2, 139–149
https://doi.org/10.1080/19466315.2017.1369899

Taylor & Francis
Taylor & Francis Group

Hazard Ratio Estimation in Small Samples

Rengyi Xu[a], Pamela A. Shaw[a], and Devan V. Mehrotra[b]

[a]Department of Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA; [b]Biostatistics and Research Decision Sciences, Merck Research Laboratories, North Wales, PA

ABSTRACT

When comparing survival times between groups in the setting of proportional hazards, the Cox model is typically used for estimation and inference, the latter based on large sample considerations. Mehrotra and Roth introduced a generalized log-rank (GLR) method for better statistical efficiency in estimating relative risk in small samples. In this article, we propose a refined GLR (RGLR) statistic by eliminating an unnecessary approximation in the development of the original GLR approach, and provide further insights into the performance of GLR and RGLR statistics. We also extend RGLR to allow for tied event times. We show across a variety of simulated scenarios that RGLR provides a smaller bias than commonly used Cox model, parametric models and GLR in small samples (up to 40 subjects per group), and has notably better efficiency relative to Cox and parametric models in terms of mean squared error. The RGLR method also consistently delivers adequate confidence interval coverage and Type I error control, while parametric methods and the Cox model tend to under-perform on that front in small samples. We further show that while the performance of the parametric model can be significantly influenced by misspecification of the true underlying survival distribution, the RGLR approach provides a consistently low bias and high relative efficiency. We apply all competing methods to data from two clinical trials. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received September 2016
Revised March 2017

KEYWORDS
Cox regression; Generalized log-rank statistic; Relative risk; Small sample; Tied event times

1. Introduction

In a typical survival analysis comparison of two groups, the hazard ratio, often called the relative risk, is generally the focus of inference. If the hazard ratio can be assumed constant throughout time, that is, if the two groups have proportional hazard functions, it is conventional to use the Cox proportional hazards model for estimation of relative risk and the log-rank test for hypothesis testing; the latter can be derived as a score test via the Cox partial likelihood function (Cox 1972). However, Cox regression is a large-sample method and small-sample sizes (10–100 subjects per group) are quite common in real-data applications such as early-phase clinical trials (Pocock 1983). Besides randomized clinical trials, observational studies involving a rare disease also often have limited sample sizes. Therefore, it is important to study analysis methods for failure time data in small samples. Johnson et al. (1982) performed a simulation study to investigate the Cox model with one binary indicator as the covariate under small samples. They found that when total sample size exceeds 40, there is no censoring, and there are equal number of subjects in the two groups, the bias of the estimated log hazard ratio is reasonably low and the sample variance is similar to the asymptotic variance. However, in smaller samples, there are nontrivial differences between the actual and asymptotic formula-based variances.

To improve the estimation and inference of relative risk in studies with small sample sizes, Mehrotra and Roth (2001) proposed a method based on a generalized log-rank (GLR) statistic for the two-group comparison. They showed that even though asymptotically the GLR method has similar performance to the Cox approach, when the sample size is small, GLR is notably more efficient than the Cox approach, in terms of mean squared error (MSE) for the log relative risk when there are no ties.

In this article, we refine the GLR method by replacing previously formulated "approximate" nuisance parameters with "exact" counterparts, for settings with and without tied event times. We show through numerical studies that the refined GLR (RGLR) statistic provides a notably smaller bias than the GLR statistic and more commonly used methods such as the Cox and parametric models, while providing a high relative efficiency and maintaining coverage for 95% confidence intervals. We provide further insights into the GLR statistic by developing an alternate estimation approach for the nuisance parameters. We also compare the performance of the RGLR statistic to parametric models, the Cox model, and the GLR approach. Furthermore, we examine RGLR's performance with respect to Type I error and confidence interval coverage, and we compare RGLR with correctly and incorrectly specified parametric models. Section 2 includes the derivation of RGLR statistic for testing and estimation, where we also describe a different approach for estimating the nuisance parameters. In Section 3, we study the numerical performance of the competing methods through a simulation study. In Section 4, we apply the different methods to data from two real clinical trials, and end with concluding remarks in Section 5.

CONTACT Rengyi Xu rengyi.xu@takeda.com Department of Epidemiology and Biostatistics, University of Pennsylvania, Philadelphia, PA 19104.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/SBR.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/SBR.

2. Methods

2.1. Refined GLR Statistic for Hypothesis Testing with No Tied Event Times

Suppose there are two treatment groups A and B, and we randomize N_A and N_B subjects to each of the groups, respectively. We assume for now that there are no tied observations. Let $t_1 < t_2 < \cdots < t_k$ denote the ordered observed event times for the combined data. Let T denote the random variable for the event time, and $S_B(t)$ and $h_B(t)$ denote the survival and hazard function for T in group B. By definition, we can write $S_B(t) = P(T > t) = \exp(-\int_0^t h_B(x)\mathrm{d}x)$, so that

$$\begin{aligned} P(t_{i-1} < T \le t_i | T > t_{i-1}) &= 1 - P(T > t_i | T > t_{i-1}) \\ &= 1 - \exp(-p_i), \end{aligned} \tag{1}$$

where $p_i = \int_{t_{i-1}}^{t_i} h_B(x)\mathrm{d}x$. In the development of the original GLR statistic, $1 - \exp(-p_i)$ was simplified to p_i by invoking a first-order Taylor series approximation (Mehrotra and Roth 2001). In this article, motivated by a desire to reduce bias, we use the exact value of $1 - \exp(-p_i)$ in a refined GLR statistic (RGLR).

Let the random variables D_{iA}, D_{iB} denote the number of events in group A and B at t_i, respectively, and let $D_i = D_{iA} + D_{iB}$. Let the random variables R_{iA}, R_{iB} denote the number of subjects still at risk at time t_i in group A and B, respectively. We then let r_{iA} and r_{iB} denote the observed number of subjects at risk at time t_i in group A and group B, respectively, and the observed total number of events and observed total number of subjects at risk at time t_i as d_i and r_i, respectively. At t_i, we can think of D_{iB} as following a binomial distribution with probability $\pi_{iB} = 1 - \exp(-p_i)$ and r_{iB} trials. Then, under the proportional hazards assumption, it follows that the number of events in group A, D_{iA}, will follow a binomial distribution with probability $\pi_{iA} = 1 - \exp(-\theta p_i)$ and r_{iA} number of trials, where θ is the hazard ratio for group A versus B. Let $G_i = \{j : \max(0, d_i - r_{iB}) \le j \le \min(d_i, r_{iA})\}$. Given $d_i, r_{iA}, r_{iB}, p_i, \theta$, the conditional distribution of D_{iA} follows a non-central hypergeometric distribution, and we can write the probability function as

$$\lambda_{iA} \equiv P(D_{iA} = d_{iA} | R_{iA} = r_{iA}, R_{iB} = r_{iB}, D_i = d_i, p_i, \theta) \tag{2}$$

$$= \frac{\binom{r_{iA}}{d_{iA}}\binom{r_{iB}}{d_{iB}}(1-e^{-p_i})^{d_{iB}} e^{-p_i(r_{iB}-d_{iB})}(1-e^{-\theta p_i})^{d_{iA}} e^{-\theta p_i(r_{iA}-d_{iA})}}{\sum_{j\in G_i}\binom{r_{iA}}{j}\binom{r_{iB}}{d_i-j}(1-e^{-p_i})^{d_i-j} e^{-p_i(r_{iB}-d_i+j)}(1-e^{-\theta p_i})^{j} e^{-\theta p_i(r_{iA}-j)}}. \tag{3}$$

Under the assumption of $d_i = 1\ \forall i$, the conditional mean and variance of D_{iA}, denoted by $E_{iA}(r_{iA}, r_{iB}, \theta, p_i)$ and $V_{iA}(r_{iA}, r_{iB}, \theta, p_i)$, can be derived as the following expressions:

$$E_{iA}(r_{iA}, r_{iB}, \theta, p_i) = \sum_{G_i} d_{iA}\lambda_{iA} = \frac{r_{iA}(e^{\theta p_i} - 1)}{r_{iA}(e^{\theta p_i} - 1) + r_{iB}(e^{p_i} - 1)} \tag{4}$$

$$\begin{aligned} V_{iA}(r_{iA}, r_{iB}, \theta, p_i) &= \sum_{G_i} d_{iA}^2\lambda_{iA} - \left(\sum_{G_i} d_{iA}\lambda_{iA}\right)^2 \\ &= \frac{r_{iA}(e^{\theta p_i} - 1) r_{iB}(e^{p_i} - 1)}{[r_{iA}(e^{\theta p_i} - 1) + r_{iB}(e^{p_i} - 1)]^2}. \end{aligned} \tag{5}$$

Note that the vector of nuisance parameters $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)$ is unknown and needs to be estimated. We use an unconditional approach, as suggested by Mehrotra and Roth (2001). The estimate of nuisance parameter p_i is found by maximizing the product of two unconditional binomial likelihoods, $\mathrm{Bin}(r_{iA}, 1 - \exp(-\theta p_i))$ and $\mathrm{Bin}(r_{iB}, 1 - \exp(-p_i))$:

$$L(p_i|\theta) = \pi_{iA}^{d_{iA}}(1-\pi_{iA})^{r_{iA}-d_{iA}}\pi_{iB}^{d_{iB}}(1-\pi_{iB})^{r_{iB}-d_{iB}} \tag{6}$$

$$= (1-e^{-\theta p_i})^{d_{iA}}(e^{-\theta p_i})^{r_{iA}-d_{iA}}(1-e^{-p_i})^{d_{iB}}(e^{-p_i})^{r_{iB}-d_{iB}}. \tag{7}$$

Because we are assuming no ties, the solution can be simplified to

$$\tilde{p}_{i,\theta} = \begin{cases} \log\left(\frac{\theta r_{iA}+r_{iB}}{\theta r_{iA}+r_{iB}-1}\right), & \text{when } d_{iA}=0, d_{iB}=1 \\ \frac{1}{\theta}\log\left(\frac{\theta r_{iA}+r_{iB}}{\theta r_{iA}+r_{iB}-\theta}\right), & \text{when } d_{iA}=1, d_{iB}=0. \end{cases} \tag{8}$$

Let $\tilde{\boldsymbol{p}}(\theta)$ denote the estimated nuisance parameter vector $\boldsymbol{p}$, where $\tilde{\boldsymbol{p}}(\theta) = (\tilde{p}_{1,\theta}, \tilde{p}_{2,\theta}, \ldots, \tilde{p}_{k,\theta})$. Then, the RGLR test statistic for the general null hypothesis $H_0 : \theta = \theta_0$ is

$$\mathrm{RGLR}[\theta_0, \tilde{\boldsymbol{p}}(\theta_0)] = \frac{\sum_{i=1}^k [d_{iA} - E_{iA}(r_{iA}, r_{iB}, \theta_0, \tilde{p}_{i,\theta_0})]^2}{\sum_{i=1}^k V_{iA}(r_{iA}, r_{iB}, \theta_0, \tilde{p}_{i,\theta_0})}. \tag{9}$$

The reference distribution for the RGLR statistic is approximated with an F-distribution with degrees of freedom 1 and k^*, where $k^* = \sum_i \min(d_i, r_i - d_i, r_{iA}, r_{iB})$. This is the same distribution as that used for the original GLR statistic (Mehrotra and Roth 2001). We conjecture that our RGLR statistic has the same reference distribution as the GLR statistic because we only changed the approximate nuisance parameters in the original GLR formulation with "exact" counterparts, which presumably should not affect the distribution. This is analogous to using different estimators of variance components (nuisance parameters), but the same reference null distributions in common linear mixed effects analyses. Our conjecture is strongly supported via simulations in Section 3. Note that under the most commonly used null hypothesis $\theta_0 = 1$, estimation of the nuisance parameters is no longer required, and the RGLR statistic reduces to the usual log-rank test statistic (Mantel 1966), which has an asymptotic distribution of χ_1^2.

As sample size increases, the estimate of p_i approaches zero because the time interval becomes smaller between two consecutive events and the probability of having an event in the interval approaches zero. It follows using L'Hopital's rule that when $p_i \to 0$, the RGLR statistic reduces to the score statistic from the Cox model. This demonstrates that the RGLR statistic is asymptotically similar to the Cox score statistic; this theoretical expectation is supported using simulations in Section 3.

2.2. Estimation of Nuisance Parameters

The development above is similar to the logic provided by Mehrotra and Roth (2001). However, to provide additional insight, we show that in the set up of Mehrotra and Roth's GLR statistic, the estimated nuisance parameter $\tilde{p}_{i,\theta}$ can also be estimated using the inverse-variance weighted average of the corresponding estimates of the failure probability in each group. Recall that we think of the number of events at time t_i in group

A and B as following two binomial distributions with probability π_{iA} and π_{iB}, respectively. In the setting of GLR, $\pi_{iB} = p_i$ and $\pi_{iA} = \theta p_i$ using the Taylor approximation. Therefore, there are two natural estimates of the failure probability, $\hat{\pi}_{iB} = d_{iB}/r_{iB}$ from group B and $\hat{\pi}_{iA} = d_{iA}/r_{iA}$ from group A. Thus, we have two estimates of the nuisance parameter, namely $\hat{p}_{iB,\theta} = d_{iB}/r_{iB}$ and $\hat{p}_{iA,\theta} = d_{iA}/(\theta r_{iA})$. Hence,

$$\text{var}(p_{iA}|R_{iA} = r_{iA}) = \frac{\text{var}(D_{iA})}{\theta^2 r_{iA}^2} = \frac{p_i(1-\theta p_i)}{\theta r_{iA}} \quad \text{and}$$
$$\text{var}(p_{iB}|R_{iB} = r_{iB}) = \frac{\text{var}(D_{iB})}{r_{iB}^2} = \frac{p_i(1-p_i)}{r_{iB}}.$$

Accordingly, if we equate p_i with the inverse-variance weighted average of $\hat{p}_{iA,\theta}$ and $\hat{p}_{iB,\theta}$, that is, set

$$p_i = \frac{\frac{\hat{p}_{iA,\theta}}{\text{var}(p_{iA}|R_{iA}=r_{iA})} + \frac{\hat{p}_{iB,\theta}}{\text{var}(p_{iB}|R_{iB}=r_{iB})}}{\frac{1}{\text{var}(p_{iA}|R_{iA}=r_{iA})} + \frac{1}{\text{var}(p_{iB}|R_{iB}=r_{iB})}},$$

and solve for p_i, we get the same estimated $\tilde{p}_{i,\theta}$ as that obtained via maximization of the product of the aforementioned two Binomial distributions (direct MLE approach). Since the formula for $\tilde{p}_{i,\theta}$ is somewhat complex, using the inverse-variance weighted average approach provides an intuitive and simple path to estimate the nuisance parameters.

In the setting of RGLR, however, these two approaches do not give the same estimates, because the relationship between the nuisance parameter p_i and failure probability π_{iB} is no longer linear. There are no simple closed-form solutions for the nuisance parameters using the inverse-variance weighted average approach. Although numerical solutions can still be achieved, we prefer the direct MLE approach because it delivers an exact closed-form solution. Further details of the derivation using the two approaches for the RGLR statistic can be found in the supplementary materials.

2.3. Inference using the Refined GLR Estimator for Relative Risk

The RGLR statistic is in quadratic form, which guarantees a unique minimum. Because small values of the RGLR statistic support the null hypothesis, we can derive an estimator for relative risk, $\hat{\theta}_{\text{RGLR}}$, by finding the θ that minimizes the RGLR test statistic.

The confidence interval of the RGLR estimator can then be calculated using $F(1, k^*)$ as the reference distribution. Therefore, the $100(1-\alpha)\%$ confidence interval for $\hat{\theta}_{\text{RGLR}}$ is $(\theta^L_{\text{RGLR}}, \theta^U_{\text{RGLR}})$, where

$$\theta^L_{\text{RGLR}} = \inf_{\theta}\{\theta : \text{RGLR}(\theta, \tilde{\boldsymbol{p}}(\theta)) \le F_\alpha(1, k^*)\} \tag{10}$$
$$\theta^U_{\text{RGLR}} = \sup_{\theta}\{\theta : \text{RGLR}(\theta, \tilde{\boldsymbol{p}}(\theta)) \le F_\alpha(1, k^*)\}. \tag{11}$$

2.4. Extension of RGLR to Accommodate Tied Event Times

In this section, we extend the RGLR statistic to allow for tied event times so that the method is more applicable for real datasets. There are several approaches for handling ties in the Cox model, including Breslow (1974), Efron (1977), and Kalbfleisch and Prentice (1973). Mehrotra and Roth (2011) extended the GLR statistic to incorporate ties following analogs of Kalbfleisch and Prentice's and Efron's approaches. We propose to use Efron's approach to handle ties for the RGLR statistic given that it is easier to implement.

With ties, the previous assumption that $d_i = 1\ \forall i$ no longer holds, and the conditional expected value and variance functions need to be updated to average over all possible untied orderings of tied event times at each time point i. Suppose in the time interval $(t_{i-1}, t_i]$, there are $d_i(>1)$ event times given by $t_{i,1} < t_{i,2} < \cdots < t_{i,d_i}$. Now, if we construct an *average* 2×2 life table at the unobserved true event time $t_{i,j}$, the average number of failure event times for group A and B is d_{iA}/d_i and d_{iB}/d_i, respectively, and the average number of subjects still at risk is $r_{iA} - jd_{iA}/d_i$ and $r_{iB} - jd_{iB}/d_i$ for group A and B, respectively, where $j = 1, 2, \ldots, d_i$. Then, summing across the d_i time points, we get

$$\bar{E}_{iA}(\theta, p_{i,1}, p_{i,2}, \ldots, p_{i,d_i}) = \sum_{j=1}^{d_i} E_{iA}\left(r_{iA} - (j-1)\frac{d_{iA}}{d_i}, r_{iB} - (j-1)\frac{d_{iB}}{d_i}, \theta, p_{i,j}\right) \tag{12}$$

$$\bar{V}_{iA}(\theta, p_{i,1}, p_{i,2}, \ldots, p_{i,d_i}) = \sum_{j=1}^{d_i} V_{iA}\left(r_{iA} - (j-1)\frac{d_{iA}}{d_i}, r_{iB} - (j-1)\frac{d_{iB}}{d_i}, \theta, p_{i,j}\right), \tag{13}$$

where E_{iA} and V_{iA} are shown in equations (4) and (5), using the margins of the *average* 2×2 life table at each of the unobserved true event times for the d_i events.

We derive the estimated nuisance parameter using the likelihood approach as before, where now

$$\begin{aligned} L(p_{i,j}|\theta) &= \pi_{iA}^{d_{iA}/d_i}(1-\pi_{iA})^{r_{iA}-jd_{iA}/d_i}\pi_{iB}^{d_{iB}/d_i} \\ &\quad \times (1-\pi_{iB})^{r_{iB}-jd_{iB}/d_i} \qquad (14) \\ &= (1-e^{-\theta p_{i,j}})^{d_{iA}/d_i}(e^{-\theta p_{i,j}})^{r_{iA}-jd_{iA}/d_i} \\ &\quad \times (1-e^{-p_{i,j}})^{d_{iB}/d_i}(e^{-p_{i,j}})^{r_{iB}-jd_{iB}/d_i}. \qquad (15) \end{aligned}$$

To find the nuisance parameter that maximizes equation (15), we take the log and the first-order derivative respect to $p_{i,j}$ and set it to zero. The estimating equation is

$$\frac{d_{iA}}{d_i}\cdot\frac{\theta e^{-\theta p_{i,j}}}{1-e^{-\theta p_{i,j}}} - \theta\left(r_{iA} - j\frac{d_{iA}}{d_i}\right) + \frac{d_{iB}}{d_i(1-e^{-p_{i,j}})} - \left(r_{iB} - j\frac{d_{iB}}{d_i}\right) = 0. \tag{16}$$

The estimating Equation (16) is a nonlinear function of $p_{i,j}$, and there is no closed-form solution. Therefore, we use a numerical approach to solve for $p_{i,j}$ at $t_{i,j}$; let $\tilde{\boldsymbol{p}}(\theta)$ denote the estimated nuisance parameter matrix, where entry (i, j) is denoted as $\tilde{p}_{i,j,\theta}$. Therefore, using Efron's approach to extend RGLR for tied event times, the RGLR^E test statistic for the null hypothesis $H_0 : \theta = \theta_0$ is

$$\text{RGLR}^E[\theta_0, \tilde{\boldsymbol{p}}(\theta_0)] = \frac{\sum_{i=1}^{k}[d_{iA} - \bar{E}_{iA}(\theta_0, \tilde{p}_{i,j,\theta_0})]^2}{\sum_{i=1}^{k}\bar{V}_{iA}(\theta_0, \tilde{p}_{i,j,\theta_0})}. \tag{17}$$

The reference distribution used for RGLR^E is an F-distribution with degrees of freedom 1 and k^*, where $k^* = \sum_i \min(d_i, r_i - d_i, r_{iA}, r_{iB})$. This is the same distribution as that used for RGLR with no ties. Again, this approximation is based on a conjecture that is strongly supported by simulations, as shown later. Of note, RGLR^E and RGLR are identical when there are no tied event times.

3. Simulation Study

We first compared the performance of the RGLR statistic to the Cox proportional hazards and parametric models, and to the GLR approach when there are no tied observations. We carried out a simulation study to examine issues of bias, efficiency, Type I error and the nominal 95% confidence interval coverage. For estimation with the parametric model, we examined estimation under the true versus a misspecified distribution for the simulated survival times.

For each of the N_A and N_B subjects in group A and B, independent entry time e_{ij} was generated from a uniform distribution on (0, T), where i indicates subject and $j = 1, 0$ indicates group A or B, respectively. Independent of the entry time, survival time s_{iA} was generated from Weibull (rate= 0.5θ, shape=2), and s_{iB} was generated from Weibull (rate $= 0.5$, shape $= 2$), so that the hazard ratio was θ. Note that the probability density function of a Weibull distribution with shape parameter α and rate parameter λ is $f(t) = \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha})$. The trial time for each subject was hence $t_{ij} = \min(s_{ij}, T - e_{ij})$.

We varied the sample size, percentage of censoring, and the hazard ratio between the two groups to compare the performance of the different methods. Sample size per group was varied as $N_A = N_B = 10, 20, 40, 100$. We considered percentage of censoring for the total sample of 0% and 50%. The percentage of censoring was controlled by changing the final analysis time T. For example, for 20 subjects per group with true log hazard ratio of 0.6, with $T = 2$ the mean censoring was 50.7% and the average number of events was 20.3. The log hazard ratio, denoted by $\ln(\theta)$, took values of 0, 0.6 and 1.2. Simulation results are based on 5000 replications.

Given the small-sample sizes, a problem referred to as "monotone likelihood" was encountered in some simulated datasets, where the highest event time in one group precedes the smallest event time in the other group (Bryson and Johnson 1981). Under this scenario, the hazard ratio estimate from the Cox model is infinite and not reliable. Therefore, we deleted

Table 1. Empirical bias, percent ratio of MSE relative to Cox model and coverage probability for 95% C.I. for $\ln(\theta) = 0, 0.6, 1.2$ based on 5000 simulations and an underlying Weibull distribution for the survival times*.

			$\ln(\theta) = 0$			$\ln(\theta) = 0.6$			$\ln(\theta) = 1.2$		
Censoring	N	Method	Bias	%RMSE	Cov	%Bias	%RMSE	Cov	%Bias	%RMSE	Cov
0%	10	Cox (Wald)	−0.000	100	94.2	8.42	100	94.7	8.30	100	96.5
		Cox (Score)	−0.000	100	[93.3]	8.42	100	[93.5]	8.30	100	94.4
		Weibull	−0.001	102	[92.7]	11.3	104	[93.1]	11.1	114	[93.5]
		GLR	−0.000	135	95.1	−6.50	134	94.8	−7.04	143	[93.3]
		RGLR	−0.000	114	95.1	1.52	114	95.2	1.49	117	95.7
	20	Cox (Wald)	0.001	100	94.6	4.36	100	94.6	3.99	100	95.0
		Cox (Score)	0.001	100	[94.2]	4.36	100	[94.0]	3.99	100	94.4
		Weibull	0.001	101	[93.4]	5.63	104	[93.7]	5.54	111	[93.9]
		GLR	0.001	119	94.9	−3.69	115	[94.0]	−3.46	116	[93.1]
		RGLR	0.001	108	94.9	0.53	108	94.7	0.42	108	94.7
	40	Cox (Wald)	−0.005	100	94.8	1.01	100	94.6	1.38	100	95.0
		Cox (Score)	−0.005	100	94.6	1.01	100	94.8	1.38	100	94.7
		Weibull	−0.005	101	94.5	1.76	104	94.7	2.19	109	94.5
		GLR	−0.005	111	95.1	−3.42	107	[94.3]	−2.45	106	[94.0]
		RGLR	−0.005	105	95.1	−1.12	104	94.8	−0.49	104	94.7
	100	Cox (Wald)	−0.001	100	95.1	0.75	100	94.6	0.75	100	95.0
		Cox (Score)	−0.001	100	95.0	0.75	100	94.6	0.75	100	94.9
		Weibull	−0.001	101	94.9	0.96	102	94.6	1.03	108	94.8
		GLR	−0.001	105	95.2	−1.23	103	94.5	−0.86	102	94.7
		RGLR	−0.001	102	95.2	−0.21	101	94.6	−0.05	102	95.0
50%	20	Cox (Wald)	−0.001	100	95.4	4.91	100	95.3	5.51	100	96.2
		Cox (Score)	−0.001	100	94.4	4.91	100	[94.3]	5.51	100	94.9
		Weibull	−0.000	99	[94.0]	7.24	104	[94.1]	7.40	108	94.5
		GLR	−0.001	123	96.1	−5.67	125	95.5	−5.51	128	94.6
		RGLR	−0.001	110	96.1	0.13	110	95.7	0.66	112	95.9
	40	Cox (Wald)	−0.005	100	95.5	1.45	100	95.5	1.95	100	95.2
		Cox (Score)	−0.005	100	95.0	1.45	100	95.2	1.95	100	94.9
		Weibull	−0.004	100	95.1	2.94	101	95.1	3.23	105	94.7
		GLR	−0.004	112	95.6	−4.21	112	95.6	−3.78	112	94.7
		RGLR	−0.004	105	95.6	−1.18	105	95.8	−0.69	106	95.3
	100	Cox (Wald)	0.001	100	95.3	0.57	100	94.9	0.81	100	95.2
		Cox (Score)	0.001	100	95.2	0.57	100	94.8	0.81	100	95.1
		Weibull	0.001	100	95.0	1.10	101	94.7	1.35	104	94.8
		GLR	0.001	105	95.5	−1.94	105	95.0	−1.66	104	94.6
		RGLR	0.001	102	95.5	−0.62	102	95.1	−0.37	103	95.1

* %RMSE = 100 × MSE of Cox/MSE of competing method. Results for 10 per group with 50% censoring are not reported due to monotone likelihood problems in more than 1% of the simulated datasets. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. *N*: sample size per group. Cov: coverage probability for 95% C.I. Cox (Wald): Cox proportional hazards model with Wald test. Cox (Score): Cox proportional hazards model with Score test. Weibull: Weibull regression. GLR: Generalized log-rank approach. RGLR: Refined GLR approach.

any simulated dataset in which this occurred, and if for a set of parameters of interest, there were more than 1% simulated datasets with a monotone likelihood, the results were not reported. For this reason, results for 10 subjects per group are not considered for scenarios with 50% censoring.

For each simulation scenario, we compare the empirical bias, relative efficiency and the empirical coverage probability for the 95% confidence interval for all scenarios considered for the parametric (Weibull) regression model, Cox model, GLR, and RGLR. The estimated log hazard ratio from fitting the Weibull regression is estimated by dividing the negative of the coefficient for the covariate Z, the group indicator, by the estimated scale parameter. The estimated log hazard ratio from the Cox model is the estimated coefficient for Z. Bias was reported for the case of $\ln(\theta) = 0$ and percentage bias, defined as 100 times the ratio of bias to the true value, was reported for $\ln(\theta) = 0.6$ and 1.2. The relative efficiency was calculated as the ratio of the MSE of the Cox model estimator and the estimator of the given competing method, that is, %RMSE $= 100 \times$ MSE of Cox/MSE of the competing method. Accordingly, %RMSE>100% indicates that the target method is more efficient than the Cox model.

3.1. Results on Estimation Without Tied Event Times

Per the results shown in Table 1, the RGLR statistic always had the smallest bias among the four methods and provided higher efficiency relative to the Cox model, even with 100 subjects per group. Compared to the parametric model, RGLR still had a higher relative efficiency in small samples (fewer than 20 subjects per group under 0% censoring and fewer than 40 subjects per group under 50% censoring). While GLR had the highest relative efficiency under small samples, it had a bigger bias than

Table 2. Empirical bias, percent ratio of MSE relative to Cox model and coverage probability for 95% C.I. for $\ln(\theta) = 0, 0.6, 1.2$ based on 5000 simulations and an underlying Gompertz distribution for the survival times*.

			$\ln(\theta) = 0$			$\ln(\theta) = 0.6$			$\ln(\theta) = 1.2$		
Censoring	*N*	Method	Bias	%RMSE	Cov	%Bias	%RMSE	Cov	%Bias	%RMSE	Cov
0%	10	Cox (Wald)	−0.000	100	[94.3]	8.98	100	95.1	7.96	100	96.8
		Cox (Score)	−0.000	100	[93.4]	8.98	100	[94.2]	7.96	100	94.5
		Gompertz	0.005	94	[92.3]	15.5	94	[93.0]	14.1	99	[93.7]
		Exp	−0.002	350	99.7	−39.8	210	98.5	−35.8	139	[93.4]
		Weibull	0.001	140	95.5	−5.10	138	95.2	−3.95	144	94.7
		GLR	0.000	135	95.2	−6.14	135	95.0	−7.36	143	[93.5]
		RGLR	0.000	114	95.2	2.04	115	95.7	1.17	117	96.0
	20	Cox (Wald)	0.003	100	94.5	4.06	100	95.0	3.68	100	95.1
		Cox (Score)	0.003	100	[94.0]	4.06	100	94.7	3.68	100	94.7
		Gompertz	0.005	97	[93.7]	7.36	98	[94.1]	6.90	102	94.6
		Exp	0.003	324	99.8	−39.6	132	96.9	−35.7	71	[81.2]
		Weibull	0.003	142	96.8	−10.3	133	95.7	−9.02	132	[94.0]
		GLR	0.003	119	94.9	−4.02	116	94.5	−3.55	116	[93.4]
		RGLR	0.003	108	94.9	0.21	107	95.1	0.37	108	94.9
	100	Cox (Wald)	0.002	100	95.0	0.94	100	94.9	0.81	100	95.1
		Cox (Score)	0.002	100	94.9	0.94	100	94.8	0.81	100	95.0
		Gompertz	0.002	100	94.9	1.51	100	94.7	1.41	102	94.7
		Exp	0.001	289	99.9	−40.3	34	[65.9]	−36.3	14	[5.4]
		Weibull	0.002	139	97.6	−14.2	95	[93.6]	−12.8	63	[83.6]
		GLR	0.002	105	95.1	−1.04	103	94.5	−0.80	103	94.7
		RGLR	0.002	102	95.1	−0.02	102	94.8	0.01	102	94.9
50%	20	Cox (Wald)	0.005	100	95.6	4.91	100	95.5	4.81	100	95.8
		Cox (Score)	0.005	100	94.8	4.91	100	94.7	4.81	100	94.9
		Gompertz	0.007	97	94.5	7.99	98	94.8	7.55	96	94.9
		Exp	0.006	142	97.9	−7.24	134	96.9	−7.71	138	95.8
		Weibull	0.006	112	95.7	2.93	112	95.7	1.95	119	95.5
		GLR	0.004	121	96.2	−4.35	122	95.6	−4.80	124	95.1
		RGLR	0.005	109	96.2	0.69	109	95.6	0.53	111	95.8
	40	Cox (Wald)	−0.010	100	95.0	0.45	100	95.2	1.14	100	95.5
		Cox (Score)	−0.010	100	94.6	0.45	100	94.8	1.14	100	95.0
		Gompertz	−0.010	99	94.5	2.22	99	94.8	2.64	97	94.6
		Exp	−0.008	138	97.3	−10.9	124	96.3	−10.4	116	[94.2]
		Weibull	−0.009	111	95.7	−2.28	109	95.5	−2.26	114	95.7
		GLR	−0.009	111	95.2	−4.41	111	95.4	−3.91	112	94.6
		RGLR	−0.010	105	95.2	−1.83	105	95.5	−1.18	105	95.3
	100	Cox (Wald)	0.000	100	95.1	0.62	100	95.0	0.69	100	94.8
		Cox (Score)	0.000	100	94.9	0.62	100	94.7	0.69	100	94.7
		Gompertz	0.000	100	94.7	1.26	99	94.7	1.40	99	94.6
		Exp	0.001	136	97.5	−11.1	112	95.4	−10.7	90	[91.4]
		Weibull	0.000	111	95.5	−3.11	108	95.1	−3.25	109	94.8
		GLR	0.000	105	95.1	−1.48	104	94.9	−1.47	104	94.4
		RGLR	0.000	102	95.1	−0.39	102	95.0	−0.34	102	94.9

* %RMSE $= 100 \times$ MSE of Cox/MSE of competing method. Results for 10 per group with 50% censoring are not reported due to monotone likelihood problems in more than 1% of the simulated datasets. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. *N*: sample size per group. Cov: coverage probability for 95% C.I. Cox (Wald): Cox proportional hazards model with Wald test. Cox (Score): Cox proportional hazards model with Score test. Weibull: Weibull regression. GLR: Generalized log-rank approach. RGLR: Refined GLR approach.

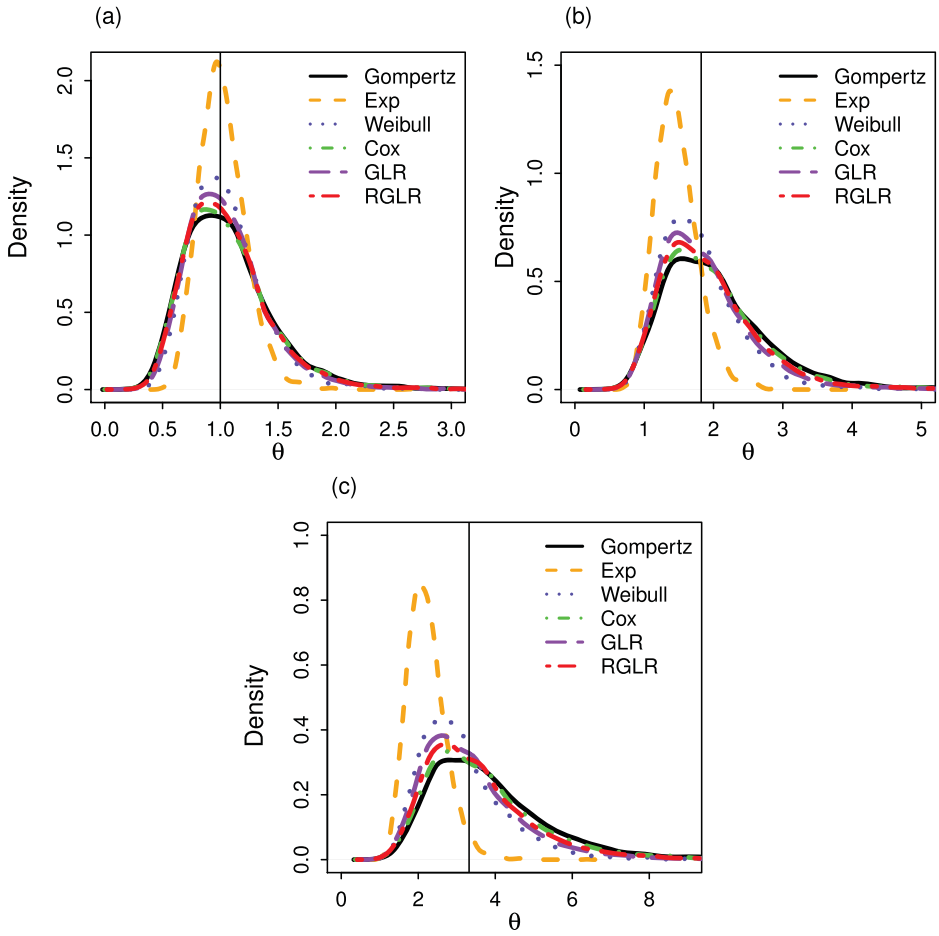

Figure 1. Empirical densities of estimators from the Gompertz, exponential, and Weibull parametric survival models, Cox model, generalized log-rank (GLR) and refined GLR (RGLR) (5000 simulations for 20 subjects per group with 0% censoring and an underlying Gompertz distribution) with a true hazard ratio of (a) 1, (b) 1.82, and (c) 3.32. A vertical line is drawn at the true hazard ratio.

RGLR and failed to maintain the nominal 95% coverage rate in some scenarios, which will be further discussed in Section 3.2. It should also be noted that the results of the parametric method are based on the true distribution. For real-data examples, it is quite difficult to make a correct assumption about the true distribution when sample size is small. When a wrong distribution is assumed, we would expect the parametric method to perform worse. Thus, the parametric method carries the risk of making the wrong assumption for the true distribution, whereas the RGLR method does not require any knowledge about the underlying distribution. We will examine the impact of misspecification of the survival distribution in Section 3.3.

3.2. Results on 95% C.I. Coverage without Tied Event Times

Table 1 also reports the empirical coverage probability for the 95% confidence interval (C.I.). Note that under the null hypothesis of $H_0 : \ln(\theta) = 0$, that is, the hazard ratio is 1, and for a two-tailed 5% significance level, 100 minus the coverage probability is equal to the Type I error rate. Therefore, a coverage probability below 95% under the null indicates an inflated Type I error. In Table 1, a value in square brackets indicates that the coverage probability is more than $Z_{0.975}$ standard errors less than the nominal rate of 95%, which implies that the Type I error rate is more than $Z_{0.975}$ standard errors above the nominal rate of 5%.

We performed a Wald test for the estimated θ using parametric (Weibull) regression and both Wald and Score tests using the Cox model. When sample size was 10 and 20 per group, the Wald test from Weibull regression and Cox Score and Cox Wald tests tended to provide an inflated Type I error rate, while our RGLR statistic controlled the Type I error rate under 5%. RGLR consistently maintained at least 95% coverage rate across all simulated scenarios. In contrast, GLR, Cox and parametric model failed to maintain the 95% coverage rate when sample size was small.

3.3. Misspecification of the Failure Time Distribution (No Tied Event Times)

As mentioned earlier, it is not always possible to assume the correct distribution when using a given parametric approach in a real-data situation. When a wrong parametric model is fit to the data, we would expect the resulting estimator to be biased. On the other hand, the RGLR approach does not make any assumption about the underlying survival distribution. We carried out a simulation study on the effect of misspecification, where the data were generated from a Gompertz distribution. The survival time in group A was generated from a Gompertz(shape = 0.5, rate = 0.2θ), and the survival time in group B was generated from a Gompertz(shape = 0.5, rate = 0.2), so that proportional hazards

Table 3. Empirical bias, percent ratio of MSE relative to Cox model and coverage probability for 95% C.I. for $\ln(\theta) = 0, 0.6, 1.2$ based on 5000 simulations and an underlying Weibull distribution for the survival times with tied observations*.

			$\ln(\theta) = 0$			$\ln(\theta) = 0.6$			$\ln(\theta) = 1.2$		
Censoring	*N*	Method	Bias	%RMSE	Cov	%Bias	%RMSE	Cov	%Bias	%RMSE	Cov
0%	10	Cox^E (Wald)	− 0.001	100	94.7	7.30	100	94.6	6.36	100	96.6
		Cox^E (Score)	− 0.001	100	[93.7]	7.30	100	[93.8]	6.36	100	94.7
		Weibull	− 0.001	99	[92.7]	11.5	101	[93.0]	11.3	107	[93.3]
		GLR^E	− 0.000	138	95.3	− 8.46	136	94.6	− 9.50	140	[92.6]
		$RGLR^E$	− 0.000	116	95.3	− 0.09	116	95.0	− 0.70	117	95.7
	20	Cox^E (Wald)	0.001	100	94.6	3.52	100	94.6	2.91	100	94.8
		Cox^E (Score)	0.001	100	[94.1]	3.52	100	[94.3]	2.91	100	94.4
		Weibull	0.002	99	[93.4]	5.69	102	[93.7]	5.62	106	[93.8]
		GLR^E	0.001	120	94.9	− 4.95	115	[94.0]	− 4.74	114	[92.7]
		$RGLR^E$	0.001	109	95.0	− 0.55	108	94.7	− 0.75	108	94.7
	100	Cox^E (Wald)	− 0.001	100	95.0	0.52	100	94.7	0.65	100	95.0
		Cox^E (Score)	− 0.001	100	95.0	0.52	100	94.6	0.65	100	95.0
		Weibull	− 0.001	101	94.9	0.96	102	94.6	1.05	108	94.8
		GLR^E	− 0.001	106	95.2	− 1.21	102	94.5	− 1.78	102	94.7
		$RGLR^E$	− 0.001	103	95.1	− 0.23	101	94.7	− 0.35	102	95.0
50%	20	Cox^E (Wald)	− 0.001	100	95.4	3.91	100	95.4	4.19	100	96.3
		Cox^E (Score)	− 0.001	100	94.5	3.91	100	94.6	4.19	100	95.1
		Weibull	− 0.000	96	[94.0]	7.45	100	[93.8]	7.67	101	94.5
		GLR^E	− 0.001	123	96.0	− 6.61	124	95.5	− 6.72	126	94.6
		$RGLR^E$	− 0.001	110	96.0	− 0.89	110	95.8	− 0.67	113	96.0
	40	Cox^E (Wald)	− 0.004	100	95.6	0.90	100	95.7	1.06	100	95.2
		Cox^E (Score)	− 0.004	100	95.2	0.90	100	95.2	1.06	100	94.8
		Weibull	− 0.004	99	94.9	3.22	99	94.7	3.41	101	94.5
		GLR^E	− 0.004	112	96.0	− 4.70	111	95.6	− 4.55	111	94.5
		$RGLR^E$	− 0.004	106	96.0	− 1.73	105	95.8	− 1.52	106	95.2
	100	Cox^E (Wald)	0.001	100	95.3	0.79	100	95.5	0.63	100	95.2
		Cox^E (Score)	0.001	100	95.1	0.79	100	95.2	0.63	100	95.1
		Weibull	0.001	100	95.0	1.93	101	95.2	1.93	101	95.2
		GLR^E	0.001	105	95.4	− 1.64	105	95.5	− 1.76	104	95.4
		$RGLR^E$	0.001	102	95.4	− 0.38	102	95.6	− 0.52	102	95.5

* %RMSE = 100 × MSE of Cox/MSE of competing method. Results for 10 per group with 50% censoring are not reported due to monotone likelihood problems in more than 1% of the simulated datasets. Coverage probability more than $Z_{0.975}$ standard errors below 95% is in square brackets. *N*: sample size per group. Cov: coverage probability for 95% C.I. Cox^E (Wald): Cox proportional hazards model using Efron's method for ties with Wald test. Cox^E (Score): Cox proportional hazards model using Efron's method for ties with Score test. Weibull: Weibull regression. GLR^E: Generalized log-rank approach using Efron's method for ties. $RGLR^E$: Refined GLR approach using Efron's method for ties.

still holds with hazard ratio θ. Each subject also had an independent entry time, and the trial was administratively censored by a fixed time T.

We again considered three different values for the log hazard ratio: $\ln(\theta) = 0, 0.6, 1.2$, percentage censoring of 0% and 50%, and varied the number of subjects per group as 10, 20, 40, 100. For each simulation, we fit the exponential, Weibull and Cox models, and applied the GLR and RGLR methods. Figure A.1 in the supplementary material shows the different hazard functions from Gompertz, Weibull, and exponential distributions.

When Gompertz was the true distribution, fitting exponential and Weibull regression under 0% censoring resulted in large bias and low percent RMSE when $\ln(\theta) > 0$, as shown in Table 2. The percentage bias from fitting exponential regression was as large as 40%, and its percent RMSE ranged from 14% to 210%. However, with a percentage bias around 30–40%, the high percent RMSE is largely meaningless. On the other hand, when the log hazard ratio was 0, exponential regression had a very small absolute bias and a high percent RMSE. However, given its poor performance in the case of non-zero log hazard ratio, this behavior indicates a tendency toward attenuation bias. Li, Klein, and Moeschberger (1996) examined the behavior of exponential regression under misspecification in the context of hypothesis testing, and found that exponential regression notably underestimates the nominal 5% alpha level when the true distribution is Gompertz and substantially overestimates when the hazard rate is decreasing. This is consistent with our finding that the exponential model performed poorly for nonzero log hazard ratio scenarios. Weibull regression, although more stable than exponential regression, still resulted in a bias of 10% or more when the sample size was at least 20 subjects per group under 0% censoring. It also started to lose efficiency as sample size increased, for example, with %RMSE=63% when $\ln(\theta) = 1.2$ under 0% censoring.

Compared to the parametric approach, the Cox model, GLR and RGLR approaches are not subject to misspecification of the underlying distribution and thus provided much more stable results. The bias of the RGLR approach was the smallest across all the simulated scenarios, and it also delivered a higher relative efficiency than the Cox model and Gompertz model when there were fewer than 100 subjects per group.

When percentage censoring increased to 50%, all methods performed better than with 0% censoring. This could be because some extreme values were censored under 50% censoring. However, exponential and Weibull regression were still the least ideal approaches. RGLR, on the other hand, consistently showed the

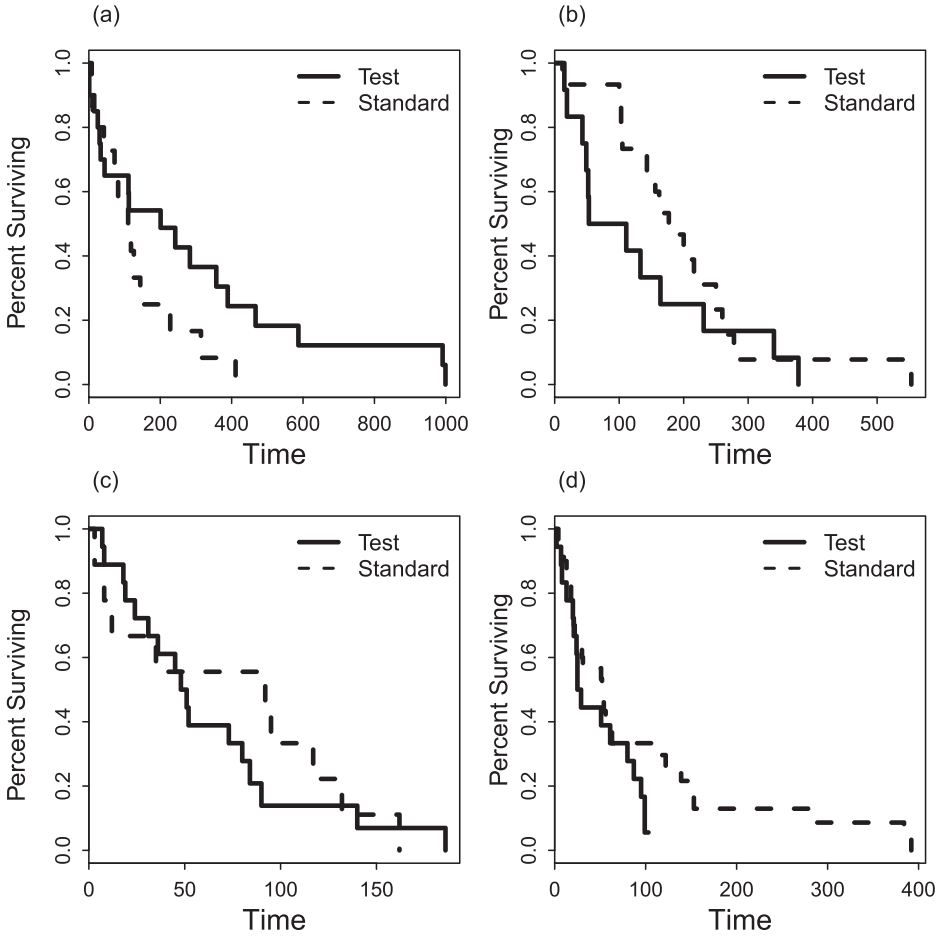

Figure 2. (a) Kaplan–Meier curves for time to death for patients with (a) squamous cell, (b) large cell, (c) adenoma cell, and (d) small cell lung cancer by treatment group. Data from Kalbfleisch and Prentice (1980).

lowest bias and high relative efficiency relative to the Cox and Gompertz models.

As shown in Table 2 and mentioned earlier, the exponential model underestimated Type I error and had poor coverage. The Cox model, especially using the score test, and GLR tended to provide a slightly lower coverage than desired. On the other hand, the RGLR approach is more stable and was able to maintain at least 95% coverage.

Figures 1(a)–(c) show the empirical densities of the estimators from the different methods with an underlying Gompertz distribution and 0% censoring and 20 subjects per group for $\ln(\theta) = 0, 0.6$, and 1.2, respectively. The vertical line is drawn at the true hazard ratio. As noted in the simulation results, in all three cases, the exponential model was adversely impacted by misspecification of the underlying true distribution. The RGLR estimates centered more closely around the true value than those from the Cox model.

3.4. Simulation Results with Tied Event Times

To compare the performance of RGLR^E to competing methods when ties in the event times are present, we again generated the data from a Weibull distribution. The set up was the same as the scenario with no tied observations, where survival time in group A was from Weibull (rate $= 0.5\theta$, shape $= 2$), and survival time in group B was from Weibull (rate $= 0.5$, shape $= 2$). Ties were created by rounding the event times to one digit after the decimal place, which is equivalent to rounding to the nearest month if the trial time unit is in years. There were approximately 15–20% tied event times, calculated as the percentage of nonunique event times in group A and B, in the the simulation studies. We compared the proposed RGLR extension for ties, RGLR^E, Weibull regression, Cox model, and GLR extension for ties, the latter two using Efron's approximation. The pattern of simulation results are very similar to that under no ties, and results are reported in Table 3.

When ties were present, with small-sample sizes, RGLR^E still delivered the smallest bias among all the methods considered, and provided higher efficiency than both Cox model that adjusts for ties using Efron's approximation and Weibull regression. It also controlled Type I error and maintained at least 95% coverage rate, while both the Cox and Weibull models tended to have inflated Type I error under small samples; of note, GLR^E failed to deliver adequate 95% confidence interval coverage in some cases.

4. Application to Two Real Datasets

We apply the RGLR and other competing methods to data from two clinical trials involving lung cancer (Kalbfleisch and Prentice 1980) and bladder cancer (Pagano and Gauvreau 2000).

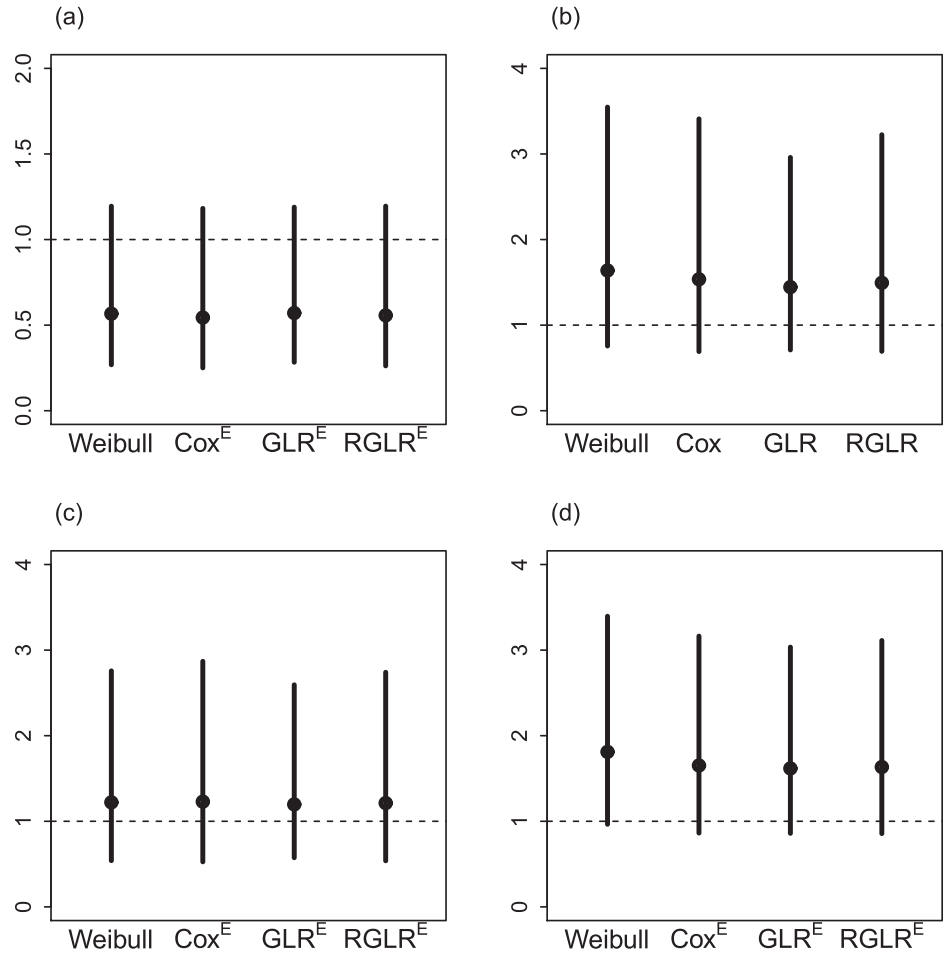

Figure 3. (a) Estimated hazard ratio and 95% confidence interval comparing test to standard chemotherapy for patients with (a) squamous cell (b) large cell (c) adenoma cell and (d) small cell lung cancer using four different methods. Cox: Cox regression. Weibull: Weibull regression. GLR: Generalized log-rank approach. RGLR: Refined GLR approach. Cox^E: Cox regression using Efron's method to adjust for tied events. Weibull: Weibull regression. GLR^E: Generalized log-rank approach using Efron's method to adjust for tied events. RGLR^E: Refined GLR approach using Efron's method to adjust for tied events. Data from Kalbfleisch and Prentice (1980).

4.1. Lung Cancer Clinical Trial

Kalbfleisch and Prentice (1980) reported results for a lung cancer trial with 137 male patients. There were 69 patients randomized patients to a standard chemotherapy and 68 patients to a test chemotherapy. Patients were categorized into four histological tumor types: squamous, small cell, adenoma, and large cell. The outcome variable was time to death (in days). Kaplan–Meier curves comparing patients on standard and test chemotherapy with different cell types are presented in Figure 2.

There were no tied event times in the large cell group, so we applied Weibull regression, Cox model, GLR and RGLR. The remaining groups all had some tied event times; therefore, we applied Weibull regression, Cox model with Efron's approximation for ties, GLR^E, RGLR^E. For patients with large cell group, GLR and RGLR provided a smaller estimated hazard ratio (test/standard) and narrower 95% C.I. than Weibull and Cox model, as shown in Figure 3(b). The estimated hazard ratio (95% C.I.) was 1.64 (0.76, 3.55) using Weibull regression, 1.54 (0.69, 3.41) using Cox regression, 1.44 (0.71, 2.96) using GLR and 1.49 (0.69, 3.22) using RGLR. For patients with squamous, adenoma, and small cell types, Weibull, Cox model with Efron's approximation, GLR^E and RGLR^E provided similar results, as shown in Figures 3(a), (c), and (d). The true hazard ratio is unknown in a real-data example, but based on our simulation results, the RGLR approach has the smallest bias and maintains coverage for 95% C.I. in small samples, and thus, is expected to be closer to the truth.

4.2. Bladder Cancer Clinical Trial

Pagano and Gauvreau (2000) reported results for a bladder cancer clinical trial. The study included 86 patients in total, who were assigned to either placebo or chemotherapy (Thiotepa) after surgery. The outcome of interest was time to recurrence (in months). For illustration, we further divided the subjects into two groups according to the number of tumors removed at surgery, one or multiple, and assessed the treatment effect. Among patients with one tumor removed, 26 patients were on placebo and 23 were on chemotherapy. Among those with multiple tumors removed, 22 patients were on placebo and 15 were on chemotherapy. Figures 4(a) and (b) present the Kaplan–Meier curves comparing patients on placebo and chemotherapy with one or multiple tumors removed.

Because of the tied event times in the dataset, in addition to Weibull regression, we used Cox model with Efron's approximation for ties, GLR^E and RGLR^E. The four methods provided

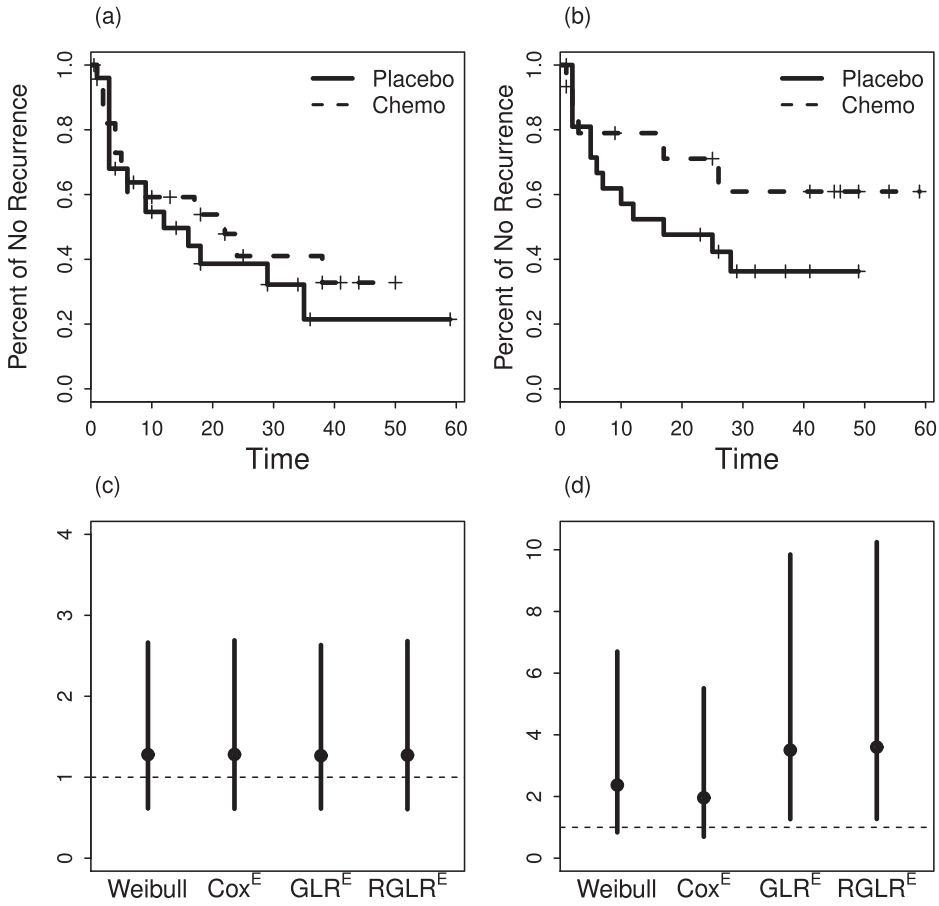

Figure 4. Kaplan–Meier survival curves for time to recurrence for bladder cancer patients with (a) one tumor and (b) multiple tumors removed at surgery by treatment group. Estimated hazard ratio and 95% confidence interval comparing placebo to chemotherapy for bladder cancer patients with (c) one tumor and (d) multiple tumors removed at surgery using four different methods. Cox^E: Cox regression using Efron's method to adjust for tied events. Weibull: Weibull regression. GLR^E: Generalized log-rank approach using Efron's method to adjust for tied events. RGLR^E: Refined GLR approach using Efron's method to adjust for tied events. Data from Pagano and Gauvreau (2000).

similar results among patients with one tumor removed, but quite different results for those with multiple tumors removed. For patients with only one tumor removed, the estimated hazard ratio (95% C.I.) of recurrence (placebo/chemotherapy) was 1.28 (0.62, 2.66) using Weibull regression, 1.28 (0.61, 2.69) using the Cox model with Efron's approximation, 1.27 (0.61, 2.63) using GLR^E and 1.27 (0.61, 2.68) using RGLR^E. For those with multiple tumors removed, the corresponding results were 2.37 (0.84, 6.70) using Weibull regression, 1.96 (0.70, 5.51) using Cox model with Efron's approximation, 3.50 (1.27, 9.85) using GLR^E and 3.60 (1.27, 10.25) using RGLR^E. As shown in Figure 4(d), both GLR^E and RGLR^E provided statistical evidence of a treatment difference based on the C.I. excluding one, while Weibull regression and Cox model with Efron's approximation for ties did not.

While Weibull and Cox regressions generated a narrower confidence interval, both of the methods tend to have inflated Type I error and lower coverage probability for 95% C.I. in small samples, as shown in our simulation studies (Section 3). Therefore, our numerical results suggest RGLR^E is expected to be closer to the truth in this example.

Of note, in both real-data examples, the estimated HR for GLR and GLR^E was always closer to one than that for RGLR and RGLR^E. This is consistent with the simulation results in Section 3 which showed that GLR and GLR^E tend to underestimate true hazard ratios that are greater than one (and, by analogy, overestimate true hazard ratios that are less than one).

5. Conclusions

Small-sample studies of time-to-event outcomes are quite common in early-phase clinical trials and observational studies of rare diseases. Thus, it is important to have methods that provide efficient hazard ratio estimation, control Type I error and maintain confidence interval coverage in small-sample settings. In this research, we developed the RGLR statistic and extended the method to allow for ties. RGLR reduces bias while maintaining high relative efficiency versus the Cox model by eliminating an unnecessary approximation in the GLR statistic. We also provided a more intuitive development using inverse-variance weighting to estimate the nuisance parameters for GLR. In addition, we demonstrated control of Type I error rate and 95% C.I. coverage in small samples for RGLR and explored the effect of misspecification of the underlying distribution on parametric models. Through simulation studies, we showed that the RGLR approach provides smaller bias relative to GLR as well as the Cox and true parametric models when the sample size per group is around 40 or less and comparable performance for larger samples. RGLR was able to consistently keep the Type I error at or below the 5% nominal level in extensive simulations, while the parametric and Cox models were observed to have an inflated

Type I error rate in small samples. Furthermore, in real-data applications, it is often challenging to know the true underlying distribution. We illustrated through simulations that when an incorrect distribution is used by a parametric regression, it can result in large bias for the estimated hazard ratio. On the other hand, the RGLR approach does not require any assumption about the true distribution and consistently delivers a very low bias with better efficiency relative to the Cox model. We recommend the use of RGLR in the setting of two-group comparisons with survival outcomes in small samples over the commonly used Cox and parametric models.

Supplementary Material

Appendix A and B: The supplementary material contains two parts. Appendix A has details showing two approaches for estimating the nuisance parameters for the RGLR statistic, and Appendix B includes a figure showing the different shapes of hazard functions of Gompertz, Weibull, and Exponential distributions. (PDF file)

References

Breslow, N. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89–99. [141]

Bryson, M. C., and Johnson, M. E. (1981), "The Incidence of Monotone Likelihood in the Cox Model," *Technometrics*, 23, 381–383. [142]

Cox, D. R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society*, Series B, 34, 187–220. [139]

Efron, B. (1977), "The Efficiency of Cox's Likelihood Function for Censored Data," *Journal of the American Statistical Association*, 72, 557–565. [141]

Johnson, M. E., Tolley, H. D., Bryson, M. C., and Goldman, A. S. (1982), "Covariate Analysis of Survival Data: A Small-Sample Study of Cox's Model," *Biometrics*, 38, 685–698. [139]

Kalbfleisch, J. D. (1980), *The Statistical Analysis of Failure Time Data*, Hoboken, NJ: Wiley. [146,147]

Kalbfleisch, J. D., and Prentice, R. L. (1973), "Marginal Likelihoods Based on Cox's Regression and Life Model," *Biometrika*, 60, 267–278. [141]

Li, Y.-H., Klein, J. P., and Moeschberger, M. (1996), "Effects of Model Misspecification in Estimating Covariate Effects in Survival Analysis for Small Sample Sizes," *Computational Statistics & Data Analysis*, 22, 177–192. [145]

Mantel, N. (1966), "Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration," *Cancer Chemotherapy Reports. Part 1*, 50, 163–170. [140]

Mehrotra, D. V., and Roth, A. J. (2001), "Relative Risk Estimation and Inference using a Generalized Logrank Statistic," *Statistics in Medicine*, 20, 2099–2113. [139,140]

——— (2011), "Improved Hazard Ratio Estimation with Tied Event Times in Small Trials," *Statistics in Biopharmaceutical Research*, 3, 456–462. [141]

Pagano, M., and Gauvreau, K. (2000), *Principles of Biostatistics*, Pacific Grove, CA: Duxbury. [146,147]

Pocock, S. J. (1983), *Clinical Trials: A Practical Approach*, Chichester, West Sussex, England: Wiley. [139]