Published in final edited form as:

Biometrics. 2017 June ; 73(2): 687–695. doi:10.1111/biom.12590.

Testing Violations of the Exponential Assumption in Cancer Clinical Trials with Survival Endpoints

Gang Han¹, Michael J. Schell^{2,3}, Heping Zhang⁴, Daniel Zelterman⁴, Lajos Pusztai⁵, Kerin Adelson⁵, and Christos Hatzis⁵

¹Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M University, 212 Adriance Lab Road, College Station, TX 77843

²The Biostatistics and Bioinformatics Department, H. Lee Moffitt Cancer Center & Research Institute, 12902 Magnolia Drive, Tampa, FL, 33612

³Oncologic Sciences, University of South Florida, 4202 E. Fowler Ave Tampa, FL, 33620

⁴Department of Biostatistics, Yale University School of Public Health, 60 College Street, New Haven, CT 06520

⁵Yale Comprehensive Cancer Center, Yale School of Medicine, 333 Cedar Street, New Haven, CT 06520

Summary

Personalized cancer therapy requires clinical trials with smaller sample sizes compared to trials involving unselected populations that have not been divided into biomarker subgroups. The use of exponential survival modeling for survival endpoints has the potential of gaining thirty five percent efficiency or saving twenty eight percent required sample size (Miller 1983), making personalized therapy trials more feasible. However, the use of exponential survival has not been fully accepted in cancer research practice due to uncertainty about whether or not the exponential assumption holds. We propose a test for identifying violations of the exponential assumption using a reduced piecewise exponential approach. Compared with an alternative goodness-of-fit test, which suffers from inflation of type I error rate under various censoring mechanisms, the proposed test maintains the correct type I error rate. We conduct power analysis using simulated data based on different types of cancer survival distribution in the SEER registry database, and demonstrate the implementation of this approach in existing cancer clinical trials.

Correspondence to: Gang Han, Department of Epidemiology and Biostatistics, School of Public Health, Texas A&M University, ghan@sph.tamhsc.edu.

The first two authors, Drs. Han and Schell, contributed equally to the work as primary authors.

Supplementary materials

Web Appendices, Tables, and Figures referenced in Sections 3.1 and 3.2, are available with this paper at the *Biometrics* website on Wiley Online Library. Specifically, in Appendix B we describe the adjustment of the H-P test for unknown λ , which is referenced in Section 3.1. In Appendix C we compare the test power of RPEXT with that of the H-P in the settings of Section 3.2. In Appendix D we demonstrate the required number of events from RPEXT under four scenarios where the simulated time-to-event data are censored, which is referenced in Section 3.2. The software program is available at the *Biometrics* website.

Keywords

Censoring; change-point modeling; failure rate; survival analysis; uniformly most powerful unbiased test

1. Introduction

Development of cancer therapies is increasingly taking into account molecular information about the tumors eventually leading to personalization of cancer therapy. Clinical trials for personalized cancer therapies typically involve fragmentation of common disease entities into multiple molecularly uniform subtypes. For example, only a subset of estrogen receptor (ER) positive patients derive benefit from chemotherapy, and thus stratifying patients by molecular subtype is necessary in the treatment for ER-positive breast cancer (Pusztai et al. 2008). In non-small cell lung cancer (NSCLC), the targeted therapy crizotinib is given only to EGFR mutation positive patients who constitute 13–15% of the NSCLC patient population (Kwak et al. 2010). As a result, there has been an increasing need for statistical methods to design and analyze personalized cancer therapy trials with small sample sizes while maintaining high estimation efficiency and test power.

Survival endpoints are clinically relevant for cancer treatment and have been used widely in clinical trials of all phases. In this article we focus on the analysis of survival distributions from single or multi-arm clinical trials instead of multivariable analysis (e.g., Cox proportional hazards model). The Kaplan-Meier estimator (KME) summarizes survival distributions by the proportion of patients surviving at each event time and typically has lower estimation efficiency and test power compared with parametric survival models that make use of all the event and censoring times to estimate the survival probability. Among parametric survival models, the exponential model is most commonly used in practice because of its simplicity and mathematical tractability, and also because it is more efficient and powerful than its nonparametric alternatives when properly used. Miller (1983) showed that without censoring, the asymptotic efficiency of the KME for estimating median survival is only 48% of the exponential approach. Meier et al. (2004) confirmed that the exponential estimate can routinely save about one-third of the patients to achieve the same accuracy for estimating median survival. However, improved efficiency is realized only if the exponential distribution assumption is valid. Motivated by this, we propose a novel statistical method for testing the validity of the exponential assumption for survival endpoints. In oncological studies, the proposed method can be used to identify significant violation of the exponential assumption. If the test indicates no violation of the exponential assumption, then one may choose to use the exponential model and take advantage of its efficiency and power. Otherwise this test will indicate up to what time the exponential assumption reasonably holds. As shown in the examples in Section 4, using such information could improve the estimation efficiency, test power, and understanding of the patient survival after receiving a treatment.

Although several tests for the exponential assumption are available, most address only uncensored data. Existing statistical approaches to testing exponentiality allowing for

censoring are primarily goodness-of-fit tests derived from the rank-based test of two samples of independent observations (Efron 1967). For example, Hollander and Proschan (1979) extended the test from Efron (1967) to compare the empirical distribution with a completely specified distribution. Hollander and Peña (1992) and Li and Doss (1993) developed Pearson type statistics by partitioning the data into different cells. Due to the space limit we do not provide a comprehensive list of existing publications on testing exponentiality, but they can be found in review papers including Koziol (1980) and Henze and Meintanis (2005). Among the tests for exponential assumption allowing censored data, the goodness-of-fit test proposed in Hollander and Proschan (1979) is by far the most commonly used: By 2014, *the Web of Science citation counts* indicated that that H-P test has been cited 61 times from all journals and cited 17 times excluding statistics and probability journals (a measure of applied use), which is the most highly cited paper for testing exponentiality, and has about three times more citation (in applied use) compared with the second highly cited paper (Akritas 1988) (cited 38 times from all journals and 5 times excluding statistics and probability journals).

The aforementioned goodness-of-fit tests for censored data have three limitations that may prohibit their use in clinical and translational research: 1, the type I error rate can be inflated if the censoring proportion is relatively high; 2, goodness-of-fit tests can not indicate whether the survival distribution is exponential up to a certain time point, which is valuable information in clinical trial practice; 3, goodness-of-fit tests are asymptotic so that they can not be used reliably in studies with small sample sizes. To overcome these limitations, we propose to test the exponential assumption by identifying significant change-points in the failure rate using an exact test (Han et al. 2012) and a backward elimination procedure to identify significant failure rate changes (Han et al. 2014). In Section 2 we introduce the proposed method and discuss the test statistic at each iteration when some of the observed times are censored. Section 3 contains two simulation studies quantifying type I error rate and test power, respectively. In Section 4, we illustrate this approach in single arm and multi-arm trials to show how it can improve the estimation efficiency, test power, and understanding of patient survival. Concluding remarks are given in Section 5.

2. Reduced Piecewise Exponential Testing of Exponentiality

Testing exponentiality is equivalent to testing whether the failure rate has statistically significant change-point(s) because the exponential assumption is valid if and only if the failure rate is a constant. We adopt the framework of reduced piecewise exponential estimate (RPEXE) in Han et al. (2014) to test exponentiality. Similar to RPEXE, the proposed test has three components: 1, a likelihood ratio test for determining the significance of any individual change-point; 2, a backward elimination procedure for identifying an arbitrary number of change-points; 3, an optional order restriction. We refer to the proposed test as *the reduced piecewise exponential test (RPEXT)*. Specifically, suppose the total sample size is N . Among the N subjects, D failed at time points $\{t_1, t_2, \dots, t_D\}$ while the other $(N - D)$ subjects were censored at $\{t_{D+1}, \dots, t_N\}$. *The total time on test (TTOT)* between two time points $t_A < t_B$ is the sum of patients' time in the interval $(t_A, t_B]$; i.e.,

$$V(t_A, t_B) = \sum_{i=1}^N \max(0, \min(t_i, t_B) - t_A). \quad (1)$$

We let d_{AB} denote the number of events in $(t_A, t_B]$. If the event time distribution $[t]$ follows an exponential distribution, we can write the *probability density function (pdf)* of t as $f(t|\lambda) = (1/\lambda) \exp\{-t/\lambda\}$ for $t > 0$, where $\lambda > 0$ is the exponential parameter denoting the mean survival and $1/\lambda$ is the failure rate. The *maximum likelihood estimate (MLE)* of λ is the total time on test of the study period divided by the number of events (Kalbeisch and Prentice 1980), i.e., $\hat{\lambda} = V(0, \max(t_1, \dots, t_N))/D$. If the exponential survival assumption is valid only between t_A and t_B , then by the memoryless property, $\hat{\lambda}_{AB} = V(t_A, t_B)/d_{AB}$ is the estimate of the exponential parameter λ_{AB} between t_A and t_B . If we further assume that 1), the exponential assumption is valid between t_A and t_B (for $t_A < t_B$) with parameter λ_{AB} and 2), valid between t_B and t_C (for $t_B < t_C$) with λ_{BC} , then testing the exponential assumption between t_A and t_C is equivalent to testing whether t_B is a significant change-point, or whether λ_{AB} is equal to λ_{BC} . The null and alternative hypotheses for testing the equivalence of λ_{AB} and λ_{BC} are

$$H_0: \lambda_{AB} = \lambda_{BC} \text{ vs. } H_1: \lambda_{AB} \neq \lambda_{BC}. \quad (2)$$

Let $x_1 = V(t_A, t_B)$ and $x_2 = V(t_B, t_C)$. The test statistic of the likelihood ratio test, which is also the uniformly most powerful unbiased (UMPU) test (Han et al. 2012), can be written as

$$\phi(x_1, x_2) = \left(\frac{x_1}{x_1 + x_2} \right)^{d_{AB}} \left(\frac{x_2}{x_1 + x_2} \right)^{d_{BC}}, \quad (3)$$

and the test rejects H_0 if $\phi(x_1, x_2)$ is less than a critical value C_α . Under type II censoring, $V(t_A, t_B)$ follows a gamma distribution $\text{Gamma}(d_{AB}, \lambda_{AB})$ with mean $d_{AB} \times \lambda_{AB}$ and variance $d_{AB} \times \lambda_{AB}^2$ if the exponential assumption is valid between t_A and t_B . Based on this gamma distribution, the distribution of $x_1/(x_1 + x_2)$ can be derived to be $\text{Beta}(d_{AB}, d_{BC})$, and the value of C_α can be quantified using this beta distribution as described in Han et al. (2012).

The second component of the reduced piecewise exponential modeling is a backward elimination procedure (Schell and Singh 1997). We describe it briefly here but greater details can be found in Han et al. (2014). This procedure initially treats all (but the last) event times $(t_1, \dots, t_{D-1})$ as potential change-points (or change-point candidates) assuming there is no order restriction. At each iteration, the procedure implements the likelihood ratio test to compute the p-value for each change-point candidate by testing the exponential parameters of two consecutive intervals. For example, at the first iteration the p-value for t_i is based on comparing the parameters from (t_{i-1}, t_i) and (t_i, t_{i+1}) given that t_{i-1} and t_{i+1} exist. At the last iteration, the test would compare parameters from $(0, t_i)$ and (t_i, ∞) if all other

change-point candidates were eliminated. At each iteration, the largest p-value (from the most non-significant candidate) is compared with a critical threshold value corresponding to the desired type I error rate. If the largest p-value at this iteration is greater than the threshold, then the corresponding candidate time point will be eliminated from the set of remaining change-point candidates. The procedure stops if all the remaining change-point candidates have p-values smaller than the threshold value or all candidates have been eliminated, where the former indicates that the rejection of exponential assumption, and the latter indicates the exponential assumption has not been shown to be invalid. Han et al. (2014) described a simulation procedure to compute the critical threshold values (corresponding to the significance levels 0.05 and 0.1 and in their Table II) for survival data of any size. Once the critical value is set for a data set it will not change at different iterations.

The third component is an optional order restriction implemented by the pooled-adjacent-violators algorithm (PAVA) that excludes those observed event times violating a prespecified order restriction from entering the backward elimination procedure. The order restriction is a presumed trend of the failure rate, which could be non-increasing, non-decreasing, increasing then decreasing, or describing then increasing.

Different from alternative tests for exponentiality (as mentioned in Section 1), RPEXT can indicate a time up to which the exponential assumption is reasonably valid. Inference about survival characteristics based on such information may facilitate the clinical decision making, e.g., analysis of a non-small-cell lung cancer study in Section 5.1 of Han et al. (2014).

Another benefit of the proposed RPEXT approach is its robustness against different censoring proportions. We derive the distribution property of the test statistic of a single change-point, and compare the type I error rates of RPEXT and an alternative approach in a simulation study in Section 3. As mentioned above, the test statistic of the LRT for (2), i.e., $\phi(\cdot, \cdot)$, is based on the two TTOTs, $V(t_A, t_B)$ and $V(t_B, t_C)$. The RPEXT builds on the model that $V(t_A, t_B)$ follows a distribution $\text{Gamma}(d_{AB}, \lambda_{AB})$ if the survival distribution between t_A and t_B is exponential with parameter λ_{AB} . Thus we can understand the robustness of RPEXT against censoring by investigating the distribution property of $V(t_A, t_B)$ in the presence of censoring. Without loss of generality, we investigate the distribution property of the total time on test for all N subjects, where D subjects had events and $(N - D)$ were censored. Let V denote the total time on test of the N subjects. Using parameter “ δ ” to indicate the censoring status, we can write the survival data as $(t_1, \delta_1), (t_2, \delta_2), \dots, (t_N, \delta_N)$, where t_i is the observed time of the i th patient, and δ_i is the censoring status such that $\delta_i = 1$ indicates the i th patient had the event and $\delta_i = 0$ indicates the i th patient was censored. So $D = \sum_{i=1}^N \delta_i$ and $V = \sum_{i=1}^N t_i$. We let $S(\cdot)$ and $f(\cdot)$ denote the survival function and pdf of the event time, and $H(\cdot)$ and $h(\cdot)$ denote the survival function and pdf of the censoring time, respectively. As in the RPEXT, under the exponential assumption, V follows a gamma distribution with parameters D and λ , i.e., $V \sim \text{Gamma}(D, \lambda)$, where λ is the exponential parameter of the distribution for event time. The gamma distribution $\text{Gamma}(D, \lambda)$ has mean $D\lambda$ and variance $D\lambda^2$.

To derive the distribution of V under censoring, we shall assume a specific censoring distribution. Here we assume an exponential censoring model because of its mathematical tractability. Let λ_c denote the exponential parameter of the censoring distribution and $f(t, \delta)$ denote the joint probability density function of t and δ . According to Efron (1967) and Habib and Thomas (1986), the joint distribution of t and δ can be written as

$$f(t, \delta) = \{f(t)H(t)\}^{I(\delta=1)} \times \{S(t)h(t)\}^{I(\delta=0)}. \quad (4)$$

With some derivation in Appendix A, we have $t \sim \exp(\lambda p_E)$ conditional on either $\delta = 1$ or $\delta = 0$, as well as unconditional on δ . Let p_E denote the proportion of failure rate from the event, i.e.,

$$p_E = \frac{1/\lambda}{1/\lambda_c + 1/\lambda} = \frac{\lambda_c}{\lambda_c + \lambda}. \quad (5)$$

The unconditional distribution of δ is binomial $\delta \sim \text{Bin}(1, p_E)$. Under independent censoring, the exponential censoring model leads to the MLE of “ $\lambda \times p_E$ ” equal to V/N . The MLE of p_E is $\hat{p}_E = D/N$. So the MLE of λ is $\hat{\lambda} = V/D$ given the invariance property. The exact distribution of V is derived in Appendix A to be gamma with parameters N and $\lambda \times p_E$, i.e., $V \sim \text{Gamma}(N, \lambda p_E)$. If we use the MLE ($\hat{p}_E$) to replace p_E , then the estimated distribution of V is $\text{Gamma}(N, D\lambda/N)$ with mean and variance

$$E(V|\hat{p}_E) = D\lambda, \text{Var}(V|\hat{p}_E) = D^2\lambda^2/N. \quad (6)$$

Recall that in the proposed RPEXT,

$$E(V) = D\lambda, \text{Var}(V) = D\lambda^2, \quad (7)$$

according to the model assumption $V \sim \text{Gamma}(D, \lambda)$ if the exponential assumption holds. Thus, the estimated means of V are identical in both the exponential censoring and RPEXT models, indicating that the estimated λ from RPEXT is unbiased and consistent in the presence of censoring. The conditional variance in the exponential censoring model, “ $(D/N) \times D\lambda^2$ ” in (6), is smaller than that in the RPEXT model with reduction in the proportion of $(1 - D/N)$. This reduction can be interpreted as coming from censoring because without censoring, p_E will have a point mass at 1 and the two variances will be equivalent. The reduction can also be interpreted as coming from the uncertainty of $\hat{p}_E$: Instead of the point estimate, e.g., MLE $\hat{p}_E = D/N$, if we impose a prior distribution to p_E , then the variance of V will depend on the prior. For example, the proposed RPEXT and the exponential censoring model will have the same (unconditional) mean and variance of V if p_E has a beta prior with parameters D and $(N - D)$. With this prior, $E(p_E) = N/D$, which is the same as the MLE, and $\text{Var}(p_E) = D(N - D)/[N^2(N + 1)]$. So the mean and variance of the total time on test V can be derived as $E(V) = E(E(V|p_E)) = E(Np_E\lambda) = D\lambda$, and

$Var(V) = E(Var(V|p_E)) + Var(E(V|p_E)) = E(Np_E^2\lambda^2) + Var(Np_E\lambda) = D\lambda^2$, which are respectively the same as the mean and variance of V in RPEXT. As a result, the proposed RPEXT approach takes into account censoring and the associated variance. The magnitude of the variance associated with censoring can be described by a beta prior $Beta(N, N - D)$ in the event probability under exponential censoring. In Section 3.1 we numerically evaluate the robustness of RPEXT against censoring in terms of the type I error in a simulation study.

3. Simulation Examples

In this section we investigate the type I error and power of the proposed RPEXT in Sections 3.1 and 3.2, respectively.

3.1 Type I Error Simulation and Comparison

We compare RPEXT with the Hollander-Proschan (H-P) test (currently the most popular test for exponentiality with censoring as described in Section 1) for the type I error investigation. The H-P test statistic can be written as the sum of the products of the parametric survival function and the drop in the Kaplan-Meier curve at every event, i.e.,

$$C = \sum_{\delta_i=1} S_{G(t_i)} \hat{f}(t_i).$$

The statistic $C^* = \sqrt{n}(C - 0.5)/\hat{\sigma}$ asymptotically follows standard normal distribution, where the formula of $\hat{\sigma}^2$ was given in Hollander and Proschan (1979). It is worth noting that an adjustment is necessary to implement the H-P test for testing exponentiality if the value of λ is unknown, which is because the original goodness-of-fit test in Hollander and Proschan (1979) was for a completely specified exponential distribution under the null (i.e., λ is known). We provide the empirical results of this adjustment in Appendix B of the web-based supplementary materials.

In this simulation study, we generate the data from the exponential distribution with parameter $\lambda = 12$ so that the mean survival time is 12 and the failure rate is 1/12. To fully evaluate the performance under different censoring scenarios, we simulate the data with 1, no censoring; 2, uniform censoring between 24 to 48 (censoring percentage 6%); 3, uniform censoring between 6 to 24 (censoring percentage 31%); 4, right censoring at 18; and 5, right censoring at 30. Sample sizes (N) in our simulation are 30, 100, 300, and 1000, which correspond to the typical sample size in range for Phase II/III cancer trials. Under each sample size and censoring mechanism, we generate 10,000 random samples from the exponential distribution with $\lambda = 12$ and then perform the H-P and RPEXT for each of the random samples at the claimed type I error rate of 0.05 in each test. After collecting the 10,000 p-values from each test, we estimate the true type I error rate of each test empirically by counting the proportion of rejections of exponential assumption among the 10,000 simulation samples. The H-P test rejects the null if the p-value is less than 0.05. The RPEXT rejects the null if the p-value is less than the critical value calculated by $\exp(-4.2331 - 0.3938 \times \log(N))$, where the details were provided in Han et al. (2014), Table III. Similar calculation for critical values can be found in Schell and Singh (1997) for linear regression.

Table 1 lists the real (empirical) type I error rate estimates from the H-P test and the RPEXT. Both tests lead to type I error rates close to 0.05 when there is no censoring (in the row “No censoring (0)”). However, the error rate from the H-P test can be away from the desired error rate of 0.05 in the case of censoring. For example, the average empirical type I error rates from the H-P test are 0.186 and 0.626, where the censoring is uniform from 6 to 24 and censoring is at 18, respectively, with the liberalness of the test becoming worse as the sample size increases. This finding is consistent with a reminder in the original paper, Hollander and Proschan (1979) (first sentence in the last paragraph of their Section 3), that “... , we remind the reader that it is easy to exhibit situations where C will be inadequate.” Moreover, in the review paper of Kostagiolas and Bohoris (2009), although they concluded the H-P test can accommodate censoring, their simulation also indicated inflation of type I error rate similar to our simulations. For example, in Table 1 of Kostagiolas and Bohoris (2009), at sample size 30 and 70% censoring, the empirical type I error rate was 0.119 when the desired rate was 0.05. The RPEXT, however, maintains the correct type I error rate with the deviation less than 0.007 away from the desired value of 0.05. Based on this simulation study, we confirm the discussion at the end of Section 2 that RPEXT is able to take into account censoring so as to maintain the correct type I error rate.

3.2 Power Simulation

We conduct power analysis using simulated data based on different cancer survival distributions in the national registry database, the Surveillance, Epidemiology, and End Results (SEER) Program registry system, years 1995 to 2004. Types of cancer include 1, lung; 2, breast; 3, stomach; 4, pancreas; 5, colon; and 6, melanomas. We simulated data that have the same failure rate trend property as the real cancer data using the generalized gamma (GG) distribution. The density of the generalized gamma distribution at time t can be written as

$$f(t|\beta, \lambda, \sigma) = \frac{l\lambda}{\sigma\Gamma(\lambda^{-2})} \times \left[\lambda^{-2}(e^{-\beta t})^{\lambda/\sigma} \right] \lambda^{-2} \times \exp \left[-\lambda^{-2}(e^{-\beta t})^{\lambda/\sigma} \right],$$

where β is the intercept, λ is the shape parameter, and σ is the scale parameter. The survival distribution is exponential if $(\lambda, \sigma) = (1, 1)$.

We choose the GG distribution because it is the most flexible parametric model among all existing survival distributions in commercial software, e.g., SAS version 9.4. Furthermore, the failure rate's trend property (e.g., increasing or decreasing) can be represented by values (λ, σ) (Cox et al. 2007). So generating data with the estimated (λ, σ) can mimic the failure rate trend of the real scenario. Such simulation enables an evaluation of the test power of RPEXT.

Specifically, we use the SEER data extracted from the National Cancer Institute (NCI) webpage to estimate the parameters of the generalized gamma distribution. The data include local, regional, and distant cancers from 1995 to 2004, and the estimation was performed with SAS software, *proc lifereg*. Table 2 shows the estimated parameter values for cancers including lung, breast, stomach, pancreatic, colon, and melanomas. We can see that for most

of the times $\lambda \in (0, 1)$ and $\sigma \in (1, 2)$. We conduct the following simulation study to investigate the required sample size to achieve 80% test power at significance level 0.05 for identifying violations of exponentiality with $\lambda = 0.1, \dots, 0.9$, and $\sigma = 1.1, \dots, 2.1$. The range of the sample size in our simulation is 20 to 500 by 10. At each sample size we generate 5000 samples without censoring. Monotonic order restriction is imposed on RPEXT for testing each sample because it is a reasonable order restriction in a typical Phase I/II cancer trial (Han et al. 2014). The power is empirically estimated as the proportion of the rejected samples. We report the smallest sample sizes to achieve 80% power (significance level $\alpha = 0.05$) and 90% power ($\alpha = 0.1$) in Tables 3 and 4. Note that the closer (λ, σ) is to $(1, 1)$, a simulation data set is more likely to be exponential. Based on Tables 3 and 4, the sample size of 100 or less is sufficient to achieve adequate power (80% power for $\alpha = 0.05$, 90% power for $\alpha = 0.1$) in general situations where $\sigma \leq 1.3$.

The simulation study shown in Table 3 and 4 does not involve censoring and is only for RPEXT. We extend this simulation in two ways. The first extension is to compare RPEXT with H-P test using the estimated (λ, σ) for each SEER data set in Table 2. Details of this comparison are given in Appendix C of the web-based supplementary materials. We found the two tests have comparable power when there is no censoring. However, RPEXT is more reliable than H-P test given the results in Section 3.1 where the type I error rate from the H-P test could be inflated with censored observations. The second extension is to investigate the impact of censoring on the test power. Detailed results are shown in Appendix D of the web-based supplementary materials. When the censoring proportion increases to 25%, 50%, and 75%, the required number of events will increase (the required number of subjects will increase more substantially) to achieve the same test power. This finding implies that the exponential distribution is less likely to be rejected in clinical trials if the censoring proportion is higher.

4. Real Applications

4.1 A Phase II Trial of Pasireotide Long Acting Release (LAR) Intervention

We demonstrate potential improvements in the estimation efficiency using RPEXT. This multi-institutional, open-label, single-arm phase II prospective clinical trial (registered at clinicaltrials.gov with number “NCT01253161”) aimed to test the effect of pasireotide, a novel somatostatin analog with avid binding affinity to multiple somatostatin receptor subtypes. The primary endpoints were overall survival (OS) and progression-free survival (PFS) of patients with Metastatic Neuroendocrine Tumors (NETs). Among the 29 NETs patients, 19 progressed or died during the trial period. We illustrate the proposed method with the PFS data: PFS was estimated using the Kaplan-Meier and exponential methods (Figure 1). The exponential estimates of median and 2-year PFS are roughly 50% more efficient than the Kaplan-Meier estimates. Taking median survival as an example, the Kaplan-Meier estimate is 11.5 months with 95% confidence interval (CI) (3.2, 19.9), while the exponential estimate is 11.0 months with 95% CI (7.6, 16.0). The two point estimates are comparable, but the exponential model’s 95% CI is roughly 4 months shorter on each end, corresponding to a 98.7% improvement in estimation efficiency.

We use the proposed RPEXT to assess whether the exponential model is appropriate for this data. We found that the minimum unadjusted p-value in the backward elimination procedure is 0.157, which is higher than 0.010, the critical value corresponding to type I error rate of 0.1 given 19 events (Han et al. (2014), Table I). As a result, there is no significant evidence against the constant failure rate hypothesis. We conclude that use of the exponential model is acceptable given the current data, and the aforementioned benefits in the survival estimation are reliable.

4.2 Application in a Genomic Predictor of Disease Relapse-Free Survival Following Taxane-Anthracycline Chemotherapy for Invasive Breast Cancer

In this section we demonstrate the benefits of RPEXT in terms of the test power in between-group comparison and interpretation of analysis results. Hatzis et al. (2011) reported the analysis of a prospective multicenter study for triple-negative breast cancer (TNBC) patients treated with sequential taxane and anthracycline chemotherapy. They developed a new genomic predictor, herein called ACES, that involves 55 gene probe sets to predict from baseline biopsies a patient's response to chemotherapy (chemo-sensitive versus chemo-insensitive) and compared it to a 30-gene genomic predictor (pathological complete response or pCR versus no-pCR) previously developed in Hess et al. (2006) and herein designated as DLDA30. A clinically useful genomic predictor of chemo sensitivity or response should also predict better survival outcome, which is distant relapse-free survival or DRFS in this study, for the predicted chemo-sensitive group. Two types of information are important in risk stratification of TNBC patients: 1, absolute survival of patients in a predicted response group (e.g., DRFS of the predicted chemo-insensitive patients); and 2, difference in survival distribution between the two predicted response groups. It is also important to identify which genomic predictor has better performance in predicting the survival endpoints.

We test the exponential assumption for each patient group identified by the two genomic predictors in the training, validation, and combined cohorts. A significant change is claimed if the p-value is less than the critical value corresponding to the significance level of 0.05. The predicted responder group (pCR) by DLDA30 and the predicted chemo-insensitive group by ACES both had a significant failure rate change around 2.5 years. The six panels (a)–(f) in Figure 2 show that the reduced piecewise exponential estimates overlap with the corresponding Kaplan-Meier estimates grouped by different outcomes of the predictions for training (panels a and b), testing (c and d), and combined (e and f) sets.

Compared with the nonparametric alternatives, RPEXT may reveal additional useful information. First, the identification of change-points for patients with predicted pCR and chemo-insensitive status suggest that these patients experience a reduction in failure rate roughly 2.5 years after initiation of taxane-anthracycline chemotherapy according to Figure 2. Second, as a commonly used nonparametric test to compare the survival in two groups, the log-rank test assumes proportional hazards. But according to Uno et al. (2014), this assumption can be mis-specified and power to detect the mis-specification is typically low with existing methods given a moderate number of observed events. More importantly, the log-rank test can result in biased type I error estimate for unbalanced scenarios where one of

the arms has a small number of events (5 events) but the other arm has more patients. As a remedy, Kellere and Chmelevsky (1983) proposed a continuity correction procedure to alleviate the bias but also mentioned that the bias could still be present after the correction. Panels (a)–(d) in Figure 2 have unbalanced sample sizes leading to overly conservative results with the log-rank test (Kellere and Chmelevsky 1983). The exponential test, however, is free of concerns about the proportional hazards and small number of events. As a result, the reduced piecewise exponential analysis may indicate significance that is not detectable with the log-rank test. For example, log-rank test of the two arms in (d) had $p=0.04$ (Figure 3, right panel, of Hatzis et al. (2011)), but the log-rank test with Yate's continuity correction led to $p=0.06$. The significance of the difference between the two arms may still remain unclear to make a clinical decision given the two p -values from the log-rank test. Using the exponential test (Han et al. 2012), the comparison of survival between the two groups before the change-point shows high significance. The exact two-sided and one-sided likelihood ratio tests lead to $p=0.02$ and $p=0.01$, respectively. This confirms that DRFS distributions of predicted chemo-sensitive and insensitive are significantly different for ER-negative patients in 2.5 years of follow-up after receiving the treatment.

Finally, additional information becomes available from comparing the two genomic predictors in terms of their association with the survival outcome. The KME can display the survival distribution and show the significance of the association based on the log-rank test $p = 0.03$ in (b), $p < 0.01$ in (c), and $p = 0.04$ in (d). Our proposed method, however, can also estimate the failure rates and indicate the time periods during which the constant failure rate assumption is not violated. This can improve understanding of the survival characteristics for patients having a specific genomic predictor outcome. Comparing the two predictors, DLDA30 may not be associated with patient survival because the two-sided p -value from the exponential test was greater than 0.1 in (a). However, patients with chemo-sensitive prediction from the ACES predictor had significantly improved DRFS in both training and validation cohorts because the two-sided p -value is 0.001 in (c) and less than 0.03 in (d) during two and half years of follow-up. In conclusion, the ACES predictor shows significant association with DRFS where the DLDA30 predictor does not for TNBC patients receiving the taxane-anthracycline chemotherapy. Furthermore, the identified failure rate change-point around two and half years for chemo-insensitive patients may be incorporated into future trial planning and design.

5. Final Remarks

Compared with KME, proper use of the exponential estimate will lead to the following benefits: i) The clinical trial will involve a smaller number of patients or require shorter accrual period without compromising the statistical testing power; ii) Speeding up the therapy approval or rejection can make the clinical trial ethically more attractive. Such benefits are valuable especially in the personalized cancer therapy studies where the patient accrual can be difficult. Instead of suggesting using the exponential distribution routinely for survival estimation, we propose the reduced piecewise exponential test (or RPEXT) to evaluate whether the exponential fitting can provide a good approximation to the data. With the simulation and real examples in Sections 3 and 4, we claim that the proposed test can

generally maintain the correct type I error rate in the presence of censoring, has sufficient test power for typical phase II and III trials, and can be implemented in clinical studies.

The normal distribution has been routinely used in analyses of continuous measurements. A number of statistical tests have been developed to check normality. In a simulation study we confirmed a sample of 71 observations from a uniform distribution could achieve 80% power to detect the violation of normality using the Anderson-Darling test at significance level 0.05. With the same logic, RPEXT may be used to test exponentiality for a survival data set. Similar to the use of normally distributed models, the exponential model could be a reasonable approximation to the true survival distribution for the whole or part of the trial period when RPEXT shows no significant violation given adequate statistical power.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Dr. Jonathan Strosberg for allowing the use of the clinical trial data in Section 4.1, and Ms. Xiuhua Zhao for the assistance in preparing the SEER data in Section 3.2. The authors also thank Drs. Gary Rosner, Chung-Yu Huang, and Chenguang Wang for their valuable comments regarding the behavior of the proposed method in the presence of censoring. We thank the editor, associate editor, and two reviewers for their valuable comments. This research is supported in part by 1) National Cancer Institute grant P30 CA 016359-33S3, and 2) the 2016–2017 Research Enhancement and Development Initiative grant from Texas A&M University, School of Public Health.

References

- Akritis MG. Pearson-type goodness-of-fit tests: the univariate case. *Journal of the American Statistical Association*. 1988; 83(401):222–230.
- Cox C, Chu H, Schneider MF, Munoz A. Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*. 2007; 26:4352–4374. [PubMed: 17342754]
- Efron B. *Proc. 5th Berkeley Sympos. Math. Statist. Prob.* New York: Prentice-Hall; 1967. The two sample problem with censored data; 831–853.
- Habib MG, Thomas DR. Chi-square goodness-of-fit tests for randomly censored data. *The Annals of Statistics*. 1986; 14(2):759–765.
- Han G, Schell MJ, Kim J. Comparing two exponential distributions using the exact likelihood ratio test. *Statistics in Biopharmaceutical Research*. 2012; 4:348–356. [PubMed: 23814641]
- Han G, Schell MJ, Kim J. Improved Survival Modeling in Cancer Research Using a Reduced Piecewise Exponential Approach. *Statistics in Medicine*. 2014; 33:59–73. [PubMed: 23900779]
- Hatzis C, Pusztai L, Valero V, Booser DJ, Esserman L, Lluch A, Vidaurre T, Holmes F, Souchon E, Wang H, et al. A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *Jama*. 2011; 305(18):1873–1881. [PubMed: 21558518]
- Henze N, Meintanis SG. Recent and classical tests for exponentiality: a partial review with comparisons. *Metrika*. 2005; 61(1):29–45.
- Hess KR, Anderson K, Symmans WF, Valero V, Ibrahim N, Mejia JA, Booser D, Theriault RL, Buzdar AU, Dempsey PJ, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of clinical oncology*. 2006; 24(26):4236–4244. [PubMed: 16896004]
- Hollander M, Peña EA. A Chi-squared goodness-of-fit test for randomly censored data. *Journal of the American Statistical Association*. 1992; 87:358–463.

- Hollander M, Proschan F. Testing to determine the underlying distribution using randomly censored data. *Biometrics*. 1979; 35:393–401.
- Kalbeisch JD, Prentice RL. The statistical analysis of failure time data. New York: Wiley; 1980.
- Kellerer AM, Chmelevsky D. Small-sample properties of censored-data rank tests. *Biometrics*. 1983; 39:675–682.
- Kostagiolas PA, Bohoris GA. Finite sample behaviour of the Hollander-Proschan goodness of fit with reliability and maintenance data; Proceedings of the 4th World Congress on Engineering Asset Management (WCEAM 2009 Congress); 2009. 831–840.
- Kozioł JA. Goodness-of-fit tests for randomly censored data. *Biometrika*. 1980; 67:693–696.
- Kwak EL, Bang Y-J, Camidge DR, Shaw AT, Solomon B, Maki RG, Ou S-HI, Dezube BJ, Jänne PA, Costa DB, et al. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *New England Journal of Medicine*. 2010; 363(18):1693–1703. [PubMed: 20979469]
- Li G, Doss H. Generalized Pearson-Fisher chi-square goodness-of-fit tests, with applications to models with life history data. *The Annals of Statistics*. 1993:772–797.
- Meier P, Karrison T, Chappell R, Xie H. The price of Kaplan-Meier. *Journal of the American Statistical Association*. 2004; 99:890–896.
- Miller RG. What price Kaplan-Meier? *Biometrics*. 1983; 39:1077–1081. [PubMed: 6671119]
- Pusztai L, Broglio K, Andre F, Symmans WF, Hess KR, Hortobagyi GN. Effect of Molecular Disease Subsets on Disease-Free Survival in Randomized Adjuvant Chemotherapy Trials for Estrogen Receptor-Positive Breast Cancer. *Journal of Clinical Oncology*. 2008; 26(28):4679–4683. [PubMed: 18662965]
- Schell MJ, Singh B. The reduced monotonic regression method. *Journal of the American Statistical Association*. 1997; 92:123–137.
- Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, Schrag D, Takeuchi M, Uyama Y, Zhao L, Skali H, Solomon S, Jacobus S, Hughes M, Packer M, Wei L-J. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology*. 2014; 32(22):2380–2385. [PubMed: 24982461]

APPENDIX A

Maximum likelihood estimates of λ and p_E

In the exponential censoring model, the parameters are the inverse of failure rates from events (λ) and censoring (λ_c), respectively. The model parameters can also be written as (p_E, λ_t) , where $p_E = \frac{1}{\lambda} / \left(\frac{1}{\lambda_c} + \frac{1}{\lambda} \right)$ is the proportion of failure rate from events and

$\lambda_t = 1 / \left(\frac{1}{\lambda_c} + \frac{1}{\lambda} \right)$ is the inverse of the failure rates' sum. We let θ denote all the model

parameters. Suppose the observations are $(t_1, \delta_1), (t_2, \delta_2), \dots, (t_N, \delta_N)$ and we let t and δ denote the time and censoring variables, respectively. Let square bracket “[.]” denote the density conditional on θ . Following formula (6) in the main text, we have

$$[t, \delta = 1] = \frac{1}{\lambda} \exp\left(-\frac{t}{\lambda}\right) \exp\left(-\frac{t}{\lambda_c}\right) \text{ and } [t, \delta = 0] = S(t|\lambda) \times h(t|\lambda_c) = \frac{1}{\lambda_c} \exp\left(-\frac{t}{\lambda}\right) \exp\left(-\frac{t}{\lambda_c}\right). \text{ So,}$$

$$[\delta = 1] = \int_0^\infty \frac{1}{\lambda} \exp\left(-\frac{t}{\lambda}\right) \exp\left(-\frac{t}{\lambda_c}\right) dt = \lambda_c / (\lambda_c + \lambda) = p_E, \text{ and } [\delta = 0] = 1 - [\delta = 1] = 1 - p_E.$$

The conditional and unconditional distribution of t can be derived using the joint distribution $[t, \delta]$ and marginal distribution $[\delta]$, i.e.,

$$[t|\delta = 1] = [t, \delta = 1] / [\delta = 1] = \frac{1}{\lambda_t} \exp\left(-\frac{t}{\lambda_t}\right), [t|\delta = 0] = [t, \delta = 0] / [\delta = 0] = \frac{1}{\lambda_t} \exp\left(-\frac{t}{\lambda_t}\right), \text{ and and}$$

$[t] = [t|\delta = 1] \times [\delta = 1] + [t|\delta = 0] \times [\delta = 0] = \frac{1}{\lambda_t} \exp\left(-\frac{t}{\lambda_t}\right)$. As a result, the likelihood function of p_E depends only on censoring status $(\delta_1, \delta_2, \dots, \delta_N)$, and the likelihood function of λ_t depends only on event/censoring times $(t_1, t_2, \dots, t_N)$. The log-likelihoods of p_E and λ_t can be written as

$$l(p_E | \delta_1, \delta_2, \dots, \delta_N) = [\delta_1, \delta_2, \dots, \delta_N | p_E] = D \log(p_E) + (N - D) \log(1 - p_E); \quad (\text{A.1})$$

$$l(\lambda_t | t_1, t_2, \dots, t_N) = [t_1, t_2, \dots, t_N | \lambda_t] = -N \log(\lambda_t) - \frac{\sum_{i=1}^N t_i}{\lambda_t}. \quad (\text{A.2})$$

Taking derivative with respect to p_E in (A.1) and λ_t in (A.2) leads to the maximum likelihood estimates $\hat{p}_E = D/N$ and $\hat{\lambda}_t = \sum_{i=1}^N t_i / N$. So $\hat{\lambda} = \hat{\lambda}_t / \hat{p}_E = \sum_{i=1}^N t_i / D$ and $\hat{\lambda}_c = \hat{\lambda}_t / (1 - \hat{p}_E) = \sum_{i=1}^N t_i / (N - D)$ given the invariance property.

Figure 1. Kaplan-Meier estimate (solid line) and exponential estimate (dashed line) of the overall survival in trial NCT01253161. Time is in months. This figure appears in color in the electronic version of this article.

Figure 2.

Kaplan-Meier estimates and reduced piecewise exponential estimates of DRFS for patients with ER negative invasive breast cancer in (a) the training set stratified by the DLDA30 predictor, (b) the training set stratified by the ACES predictor, (c) the validation set stratified by the DLDA30 predictor, (d) the validation set stratified by the ACES predictor, (e) the combined set stratified by the DLDA30 predictor, (f) the combined set stratified by the ACES predictor. In each panel the x-axis indicates the follow-up time in years, and the y-axis indicates the survival proportion. This figure appears in color in the electronic version of this article.

Empirical estimates of type I error rates from the H-P test and RPEXT under the null hypothesis of exponentiality with the true significance level (or desired type I error rate) equal to 0.05. Survival data sets were simulated separately for 4 different sample sizes $n \in \{30, 100, 300, 1000\}$ and 4 censoring mechanisms.

Table 1

	H-P test					RPEXT				
	N=30	N=100	N=300	N=1000	Avg.	N=30	N=100	N=300	N=1000	Avg.
No censoring (0)	.047	.049	.050	.051	.049	.051	.051	.054	.043	.050
Uniform 24 to 48 (6%)	.030	.029	.033	.033	.031	.048	.054	.051	.046	.050
Uniform 6 to 24 (31%)	.041	.042	.124	.535	.186	.046	.051	.051	.050	.050
At 18 (22%)	.164	.434	.908	1.000	.626	.044	.053	.052	.050	.050
At 30 (8%)	.032	.034	.036	.081	.046	.050	.053	.052	.046	.050

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Estimated (λ , σ) from the SEER data.

	Local		Regional		Distant	
	λ	σ	λ	σ	λ	σ
Lung	0.42	1.56	0.24	1.58	0.34	1.40
Breast	2.07	0.43	2.24	0.46	0.63	1.29
Stomach	0.45	1.70	0.13	1.65	0.26	1.47
Pancreatic	-0.13	1.91	0.25	1.48	0.25	1.48
Colon	0.93	1.03	0.24	1.75	0.41	1.41
Melanomas	2.82	0.37	3.34	0.27	-0.04	1.77

Required sample sizes for detecting violations of exponential distribution with different values of (λ , σ) to achieve 80% power at significant level 0.05;

Table 3

	$\sigma = 1.1$	$\sigma = 1.2$	$\sigma = 1.3$	$\sigma = 1.4$	$\sigma = 1.5$	$\sigma = 1.6$	$\sigma = 1.7$	$\sigma = 1.8$	$\sigma = 1.9$	$\sigma = 2.0$	$\sigma = 2.1$
$\lambda = 0.1$	160	120	80	50	40	30	30*	20	20	20	20
$\lambda = 0.2$	200	140	90	60	40	40*	30	30*	20	20	20
$\lambda = 0.3$	260	160	100	70*	50*	40	30	30	20	20	20
$\lambda = 0.4$	360	200	110	70*	50	40*	30	30	20	20	20
$\lambda = 0.5$	470	230	120	70	50	40	30*	30	20	20	20
$\lambda = 0.6$	500	250	130*	80	50	40	30	30	20	20	20
$\lambda = 0.7$	500	270	130	70	50	40	30	30	20	20	20
$\lambda = 0.8$	500	260	120	70	50	40	30	30	20	20	20
$\lambda = 0.9$	500*	240	100	70	50	30	30	20	20	20	20

The star symbol “*” indicates a value (λ , σ) that corresponds to an estimate from SEER data for cancer given in Table 2.

Required sample sizes for detecting violations of exponential distribution with different values of (λ , σ) to achieve 90% power at significant level 0.1;

Table 4

	$\sigma = 1.1$	$\sigma = 1.2$	$\sigma = 1.3$	$\sigma = 1.4$	$\sigma = 1.5$	$\sigma = 1.6$	$\sigma = 1.7$	$\sigma = 1.8$	$\sigma = 1.9$	$\sigma = 2.0$	$\sigma = 2.1$
$\lambda = 0.1$	170	130	90	60	50	60	30*	30	20	20	20
$\lambda = 0.2$	220	160	100	70	50	40*	30	30*	30	20	20
$\lambda = 0.3$	260	200	110	70*	50*	40	40	30	30	20	20
$\lambda = 0.4$	360	220	120	80*	60	40*	30	30	30	20	20
$\lambda = 0.5$	500	260	140	80	60	40	40*	30	30	20	20
$\lambda = 0.6$	500	290	140*	90	60	40	30	30	30	20	20
$\lambda = 0.7$	500	290	140	80	60	40	30	30	20	20	20
$\lambda = 0.8$	500	300	130	80	60	40	30	30	20	20	20
$\lambda = 0.9$	500*	260	110	70	50	40	30	30	20	20	20

The star symbol “*” indicates a value (λ , σ) that corresponds to an estimate from SEER data for cancer given in Table 2.