Journal of Statistical Research *ISSN 0256 - 422 X*
2019, Vol. 53, No. 2, pp. 93-109

TYPE I ERROR INFLATION OF LOG-RANK TEST WITH SMALL SAMPLE SIZE: A PERMUTATION APPROACH AND SIMULATION STUDIES

ZHENG WANG*

Department of Statistics, University of Pittsburgh, Pittsburgh, PA, 15260, USA
Email: zhengwang@pitt.edu

ALICIA ZHANG

Center for Design and Analysis, Amgen, Thousand Oaks, CA, 91320, USA
Email: aliciaz@amgen.com

YUQI CHEN

Center for Design and Analysis, Amgen, Thousand Oaks, CA, 91320, USA
Email: yuqic@amgen.com

QUI TRAN

Center for Design and Analysis, Amgen, Thousand Oaks, CA, 91320, USA
Email: qtran01@amgen.com

CHRIS HOLLAND

Immunocore, Rockville, MD, 20850, USA
Email: chris.holland2@immunocore.com

SUMMARY

The log-rank test is a well-accepted nonparametric test in comparing the survival time between experimental and control group in regulatory settings. However, we have observed type I error inflation as high as 28% using the test in the simulation settings we have with even moderate sample sizes. In this paper, we explore several factors that potentially contribute to the inflation by simulation. Sample size, randomization ratio and significance levels are found to be influential factors. We propose an alternative log-rank test using an approximate permutation distribution instead of the standard normal distribution. It is shown that type I error is controlled when applying the approximate permutation test to both simple clinical trial designs and complicated group sequential designs.

Keywords and phrases: approximate permutation test; log-rank test; type I error inflation

* Corresponding author

1 Introduction

The log-rank test is a well-established test to analyze time-to-event endpoints in clinical trials. In a simple trial with a single endpoint and no interim analysis, re-validation of the log-rank test via simulation is usually not necessary. In more complicated group sequential designs with sample size re-estimation and multiple endpoints, simulation is needed to illustrate operating characteristics. One important operating characteristic is the type I error, which should be strictly controlled at a pre-determined level from a regulatory viewpoint. Our simulation shows that the log-rank type I error is inflated with sample sizes up to several hundreds of patients. Further exploration indicates that the inflation exists even in a simple design with one endpoint and no adaptation.

There has been some attention in the literature to type I error using the log-rank test. Type I error will be inflated with a small sample size and unbalanced randomization ratio (Kellerer and Chmelevsky, 1983; Latta, 1981). Tang (2014) explored randomization allocation between experimental and control groups and showed through simulation that more sample size is required to control type I error when the randomization is imbalanced. Strawderman (1997) also focused on the behavior of the log-rank test when the randomization between two groups is not balanced. They pointed out that the distribution of the log-rank statistic is skewed in this case and biased. They proposed a normalizing transformation to provide a more normally distributed test statistic. However, a systematic examination of factors contributing to type I error associated with the log-rank test is lacking. On the other hand, there are variations of the log-rank test that use different variance estimators. The most well-known version and commonly implemented in statistical software is the Mantel-Cox log rank test (Mantel, 1988; Schoenfield, 1981; Klein and Moeschberger, 2006). Brown (1984) compared this version with the Peto and Peto (1972) log-rank test, which used permutation. He observed that the Mantel-Cox log rank test tends to underestimate the variance and therefore results in an inflation of type I error, while the Peto-Peto log-rank test tends to overestimate the true variance and could possibly serve as a prudent choice when controlling type I error is required. The behavior of the Peto-Peto log-rank test is also examined in our paper.

Since the distribution of log-rank test statistic deviating from normal distribution leads to type I error inflation, and there is no closed formulation of the true distribution, we propose performing hypothesis testing using permutation distribution. The permutation test obtains the null distribution by calculating all possible values of the test statistic under rearrangement of the randomization labels. When there are too many possible orderings of the data, taking a random sample of the possible replicates is asymptotically equivalent (Dwass, 1957). Neuhaus (1988, 1993) established conditional central limit theorem for survival test statistics with permutation. It's found that studentized survival test statistic can control type I error with large sample size regardless of censoring distributions, and type I error can be guarded with small sample size and equal censoring. The studentized permutation test has been studied for weighted log-rank test with random right censoring (Brendel et al., 2014), when data fail to satisfy independent and identical distribution assumption (Janssen, 1997), and when two groups have different variances (Janssen and Pauls, 2003). The popular asymptotic normal log-rank test has been shown to be anticonservative compared with the permutation test (Heller and Venkatraman, 1996). Our simulations show that 5,000 permutations can provide a good approximation to the exact distribution of the log-rank statistic. This is referred to as approximate

permutation.

This paper is organized as follows: We first identify the factors that cause type I error inflation in a simple design, then we introduce the approximate permutation log-rank test. Further simulation results show that type I error can be controlled using the approximate permutation test in both a simple design and a more complicated group sequential design. We also explore certain aspects that may impair the performance of the permutation test, such as unequal censoring. Finally, we conclude with some discussions. Throughout the paper, a simple design refers to a design with one endpoint and no adaptation. A complicated design refers to a design with two endpoints and multiple interim analyses of futility, efficacy and sample size re-estimation (SSR). The log-rank test refers to the Mantel-Cox log-rank test unless otherwise specified.

2 Factors Considered for Type I Error Inflation in a Simple Design

We first considered various factors in a clinical trial with a simple design to examine their effects on type I error inflation. Throughout this paper, α indicates type I error rate. Overall survival (OS) is used as an example of a time-to-event endpoint. The alternative hypothesis for a one-sided test assumes survival time in the experimental group is longer than the control group.

To investigate which factors affect type I error inflation, 100,000 simulations were run for each variation of the following factors: α level, dropout rate, sample size and randomization ratio. Unless otherwise noted, the general simulation setting is as follows: 400 subjects were randomized, since a typical randomized clinical trial recruits a few hundred people. No maximum follow-up time was applied, which means subjects are observed till death or dropout. Subjects were randomized to experimental or control group with a ratio of 2:1. Under the null hypothesis of no treatment difference, survival time S for both groups was generated from an exponential distribution with a median of 12 months. Censored time D follows an exponential distribution such that rate of dropout per year is 5%. Observed event time is denoted $T = \min(S, D)$ and event status $\epsilon = I(S < D)$, where $I(\cdot)$ is an indicator function, $\epsilon = 1$ indicates death is observed and $\epsilon = 0$ indicates the subject is censored. Function implemented in R package was used to implement one-sided log-rank test at α of 0.02.

2.1 α Level

Various clinical trial designs with one or multiple endpoints may require different levels of α. Holding other parameters constant, α of 0.005, 0.01, 0.02, 0.025, and 0.05 are investigated. From Figure 1, we can observe that, given the same dropout rate, smaller levels of α tend to have larger percentage of type I error inflation, calculated as the observed type I error rate divided by α. Further, the one-sided log-rank test has larger inflation of type I error compared with two-sided log-rank test.

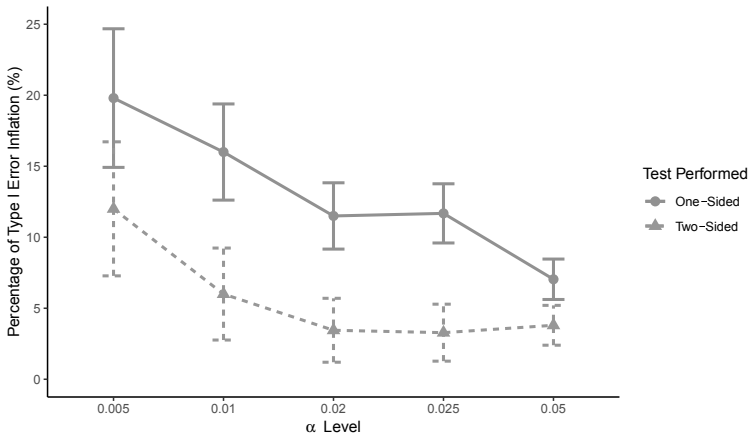

Figure 1: Type I error at different α levels with Monte-Carlo error shown as error bar

2.2 Dropout Rate

We examine the effect of subject dropout rates on type I error. Holding everything else constant, D is generated from an exponential distribution that results in the dropout rate varying from 0% to 80% per year.

From Figure 2, the inflation percentage is relatively stable even for varying dropout rates. When α is small, e.g. 0.005, there are some fluctuations, but the absolute change for different dropout rates is small.

2.3 Sample Size and Randomization Ratio between Intervention and Control Arms

Since the log-rank test assumes that the distribution of test statistic is asymptotic and Gaussian, it is of interest to examine how sample size and the randomization ratio influence type I error. In this simulation, the sample size ranges from 50 to 3,200. Although equal allocation can maximize statistical power for a given sample size, unequal allocation is often used in clinical trials to help with enrollment, particularly for promising therapies. The randomization ratio between experimental and control typically varies from 1:1 to 4:1. The simulation result is summarized in Figure 3.

From Figure 3, aligning with the fact that the log-rank test statistic is asymptotic normally distributed, there is a clear pattern that the type I error inflation is smaller when the sample size is larger and the randomization allocation between experimental and control is more balanced. However, even for a 1:1 randomization ratio, there is still type I error inflation observed when the sample size is several hundreds of patients.

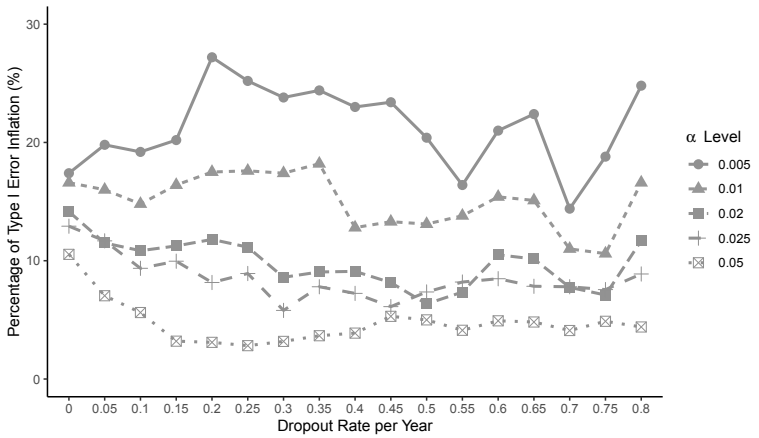

Figure 2: Type I error at different dropout rates and α levels

3 Why Log-Rank Test Has Inflated Type I Error?

As we have seen from Figure 2, type I error inflation decreases with increasing sample size. With sufficient sample sizes the log-rank test statistic approaches asymptotic normality. When the sample size is not big enough, the distribution deviates from the standard normal (Gaussian) distribution. Simulations were run to further illustrate how the distribution of log-rank test statistic is shaped.

Using the simulation setting from Section 2, 1,000,000 simulations are carried out, resulting in 1,000,000 log-rank test statistics shown in the left plot of Figure 4. The distribution density kernel and standard normal density kernel are added for comparison. The right plot of Figure 4 is the magnified version of left tail of log-rank test statistic histogram. The critical value from an asymptotic standard normal distribution is shown as the solid line in the right plot of Figure 4, and the critical value from the sampling distribution for hypothesis testing when α=0.02 is shown as the dotted line.

Using a one-sided test (percent inflation 10.13% with Monte-Carlo error 0.73% when α=0.02), the critical value to reject the null hypothesis is derived from the left tail of the standard normal distribution. From the histogram, we can see that this critical value is bigger than the critical value from the empirical log-rank test statistic's distribution, and thus leads to type I error inflation.

4 Approaches to Address Type I Error Inflation

4.1 Peto-Peto Log-Rank Test

Unlike the popular Mantel-Cox log-rank test implemented by the function in R and in SAS, the Peto-Peto log-rank test uses a permutation variance. It tends to overestimate the true variance and results in a smaller type I error rate than the pre-determined level. In light of this, the use of the

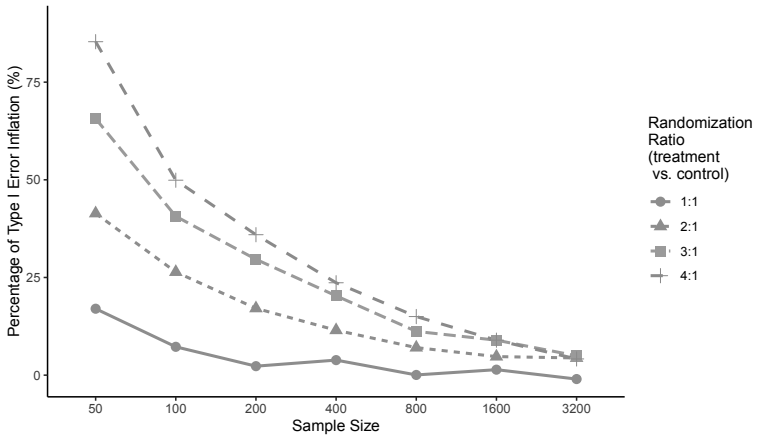

Figure 3: Type I error with different sample sizes and randomization ratios

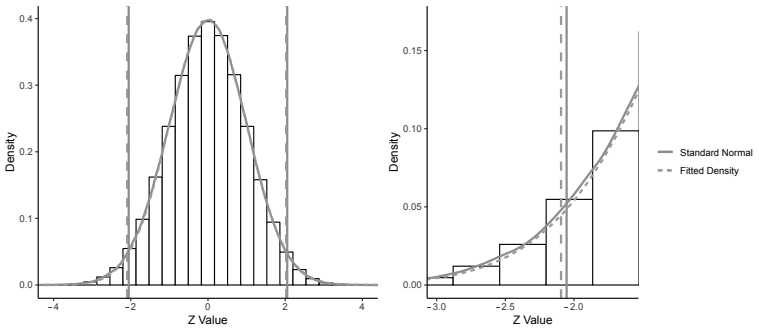

Figure 4: Log-rank test statistic distribution compared with normality

Peto-Peto log-rank test is an option to control type I error. The performance of Peto-Peto log-rank test is illustrated in the next section.

4.2 Approximate Permutation Approach

The root issue is that the distribution of the log-rank test statistic deviates from the normal distribution when the sample size is not large enough. However, in clinical trials, especially in oncology and rare diseases, it is often not feasible to increase the sample size to thousands of patients due to limited time, restricted resources, and ethical reasons. A natural solution is to find another way to correct the distribution of the log-rank test statistic. One option is the exact test. Although exact tests provide satisfying results (not shown), it is feasible from a computational standpoint only when the sample size is small, e.g. less than 50. An alternative approach to the exact test would be the

approximate permutation test (Berry et al., 2011; Davison and Hinkley, 1997). This test reduces the computation time significantly and the distribution of the test statistic may be used to have better control of the type I error rate compared to a normal approximation.

The accuracy of the permutation approximation depends on how many permutations one performs. To examine this, both normal approximation and approximate permutation tests using 1,000, 2,000 and 5,000 permutations are carried out and results from 100,000 simulation runs are shown in Figure 5. Regardless of the number of permutations, the test statistic from the original sample will serve as one of the permutations as suggested by Davison and Hinkley (1997). To perform the permutation test, we reassign the treatment group randomly such that subjects are assigned to each group according to the intended randomization ratio. The next step is to perform the log-rank test and obtain the test statistic. After repeating this process by the specified number of times, we use the distribution from the test statistics to approximate the true distribution. The p-value is calculated as the proportion of these test statistics that are smaller than the log-rank test statistic from the original sample.

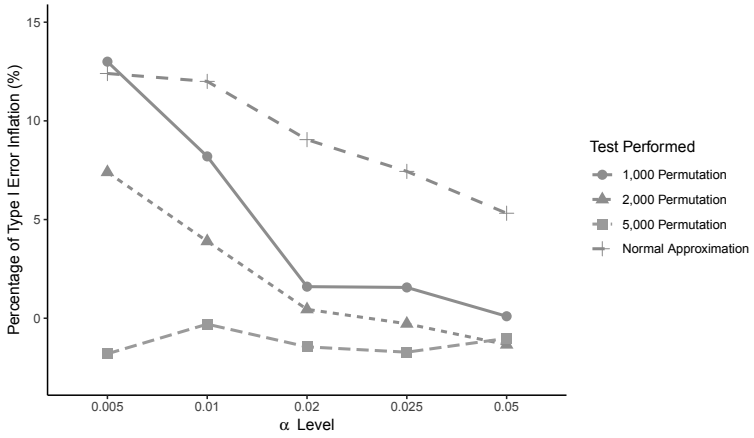

Figure 5: Performance of permutation test

Although a greater number of permutations can provide further reduction in general, the computation time is longer. From Figure 5, 5,000 permutations can control the type I error at different α levels with reasonable computation time.

Table 1 shows the type I error rates from Mantel-Cox and Peto-Peto log-rank tests, using critical values from the standard normal distribution and the permutation distribution. Five thousand permutations and $\alpha = 0.02$ was used. The Peto-Peto log-rank test is implemented in the package. It is shown in Table 1 that using the permutation test can control type I error rate for Mantel-Cox log-rank test. The Peto-Peto log-rank test results in type I error rates below pre-determined α levels, which supports the theory proposed by Brown (1984). It is interesting to note that the permutation test corrects the type I error of the Peto-Peto log-rank test by increasing the type I error level to be

Table 1: Type I error for a simple, single-test design ($\alpha = 0.02$)

Statistics	Type I error
Mantel-Cox using normal approximation	0.02181
Mantel-Cox using permutation test (5,000)	0.01971
Peto-Peto using normal approximation	0.01904
Peto-Peto using permutation test (5,000)	0.02059

Table 2: Power analysis for a simple, single-test design ($\alpha = 0.02$, power goal 80%)

Statistics	Power
Mantel-Cox using normal approximation	0.80038
Mantel-Cox using permutation test (5,000)	0.78519
Peto-Peto using normal approximation	0.77178
Peto-Peto using permutation test (5,000)	0.78432

closer to the pre-determined level.

4.3 Power Analysis

Evaluation of how power changes under the alternative hypothesis is also of interest with approximate permutation approach and asymptotic normal assumption. Here, power is examined with both Mantel-Cox and Peto-Peto log-rank tests. Similar to the simulation setup from Section 2, subjects were randomized into either the experimental or control group with a ratio of 2:1. Survival time in the experimental and control groups is generated from exponential distribution with median of 10 and 7 months respectively, so that the hazard ratio is 0.7 between experimental and control group. Censored time follows an exponential distribution with dropout rate of 5% each year. To achieve 80% power, a total of 315 subjects are assumed to be enrolled in the study. After 100,000 simulation runs, Table 2 summarizes the results under $\alpha = 0.02$.

Generally, both Mantel-Cox and Peto-Peto log-rank test with asymptotic normal approximation and approximate permutation test can achieve similar power. Peto-Peto log-rank test will cause some loss of power compared with Mantel-Cox method. Similar to what we have observed from Table 1, using approximate permutation test will increase the power of Peto-Peto log-rank test, while it will reduce the power of Mantel-Cox log-rank test.

Table 3: Type I error for a simple, single-test design with different distributions ($\alpha = 0.02$)

Method	Weibull	Lognormal
Mantel-Cox using normal approximation	0.02207	0.02141
Mantel-Cox using permutation test (5,000)	0.02017	0.01943
Peto-Peto using normal approximation	0.01807	0.01819
Peto-Peto using permutation test (5,000)	0.01997	0.01999

5 Alternative Considerations

5.1 Change of Survival Time Distribution

We have been using exponential distribution for survival time in our simulation studies. Here, other survival time distribution assumptions are considered to evaluate the generality of the permutation and asymptotic normal log-rank test performance. Four hundred subjects are to be assigned randomly into treatment and control groups with a 2:1 randomization ratio. Instead of the exponential distribution, we will consider one simulation study with survival time following a Weibull distribution (shape=2, scale=14). The other simulation study assumes survival time to follow a lognormal distribution ($\mu = 0.5$, $\sigma = 2$). Dropout rate is 5% per year. Both a 5,000 permutation test and an asymptotic test are carried out for Mantel-Cox and Peto-Peto log-rank test. Table 3 summarizes the results after 100,000 simulation runs with α = 0.02.

Comparing it with Table 1, we can observe the same feature that the type I error is inflated with a Mantel-Cox log-rank test using normal approximation. For the Peto-Peto log-rank test, the type I error is lower than α. The distribution of survival time does have an effect on type I error no matter which test is chosen, but the permutation test can correct the type I error at the pre-determined α level.

5.2 Unequal Censoring

Throughout the simulations in Section 2, 3 and 4, we have only considered independent and identically distributed time to event and equal censoring between experimental and control groups. However, this may not be necessarily true in practice. In this case, the log-rank test may not perform as well as expected. In this subsection, we will evaluate the performance of the log-rank test using asymptotic normality and an approximate permutation test when censoring is not equal between experimental and control groups.

Equal censoring may not be guaranteed in clinical trials where there are unequal dropout rates between treatment groups. This is also the case for observational studies where treatment groups are followed at different periods or under different time duration. Applying the Mantel-Cox log-rank test with asymptotic normal or permutation test while ignoring unequal censoring will cause concerns of type I error and power. In this simulation, we randomly assign 400 subjects into treatment and

Table 4: Type I error with unequal censoring between groups ($\alpha = 0.02$)

		Type I error	
Dropout rate in placebo group	Dropout rate in treatment group	Normal approximation	5,000 permutation
0.05	0.05	0.02190	0.01974
	0.10	0.07325	0.06779
	0.15	0.19376	0.18331
0.10	0.05	0.00542	0.00487
	0.10	0.02152	0.01963
	0.15	0.07119	0.06604
0.15	0.05	0.00121	0.00110
	0.10	0.00562	0.00500
	0.15	0.02190	0.01989

control groups with a 2:1 randomization ratio. Survival time for both groups follows exponential distribution with median survival equal to 12 months. Three different dropout rates per year are considered for both groups as illustrated in Table 4 after 100,000 runs of simulations.

We can see that permutation and asymptotic normal test both failed to control type I error at the pre-defined $\alpha = 0.02$ in the case of unequal censoring between treatment groups. Therefore, data should be carefully examined for informative censoring and alternative approaches should be used such as exact conditional test (Heinze et al., 2003; Heimann and Neuhaus, 1998) and nonparametric combination test (Arboretti et al., 2018).

6 Simulation Study for Group Sequential Design with Sample Size Re-estimation

We have shown that the permutation test provides an attractive alternative to control the type I error rate in a simple design where censoring is balanced. In this section, we apply this approach to a more complicated group sequential design with two primary end-points, overall survival (OS) and progression free survival (PFS). The example is an ongoing oncology clinical trial. The experiment-wise significance level is set at 0.025, consistent with the level of type-1 error allowed in clinical trials with one-sided tests. To control for multiple hypotheses testing, this significance level alpha has to be split between the two primary endpoints PFS and OS. With clinical evidence from a prior trial that there is a strong treatment effect on PFS, PFS is tested one-sided at a significance level of 0.005. In case such an effect is not demonstrated, the remaining alpha, 0.02, is reserved for testing

Table 5: Approximate timing of interim analysis for PFS and OS (Section 6.1 Setting)

Analysis	Time (Months)*	PFS (Progression Free Survival)	OS (Overall Survival)
1	19.4	Interim for futility (50% events), use $265 \times 0.5 = 133$ events	——
2	26.5	Interim for futility and sample size re-estimation (80% events), use $265 \times 0.8 = 212$ events	Interim for futility and sample size re-estimation (54% events), use $304 \times 0.54 = 164$ events
3	31.1 (Enrollment completes)	Final (265 events)	Stopping early for efficacy (69% of events), use $304 \times 0.69 = 210$ events
4	45.8	——	Final (304 deaths)

*From the date of randomization of the first subject

OS. The randomization ratio between experimental and control group is 2:1. The enrollment rate is estimated to be 13 subjects per month. There are several planned interim analyses for OS and PFS at which sample size can be re-estimated (Table 5).

6.1 Type I Error

In this trial, 394 subjects are anticipated to be enrolled with 265 PFS events and 304 OS events to achieve required power. Under the null hypothesis, OS time T_{os} and disease progression time T_{pd} for both groups is generated from the exponential distribution with a median survival time of 12 and 20 months respectively. PFS time $T_{\text{pfs}} = \min(T_{\text{os}}, T_{\text{pd}})$ with median survival time of 7.5 months.Maximum number of subjects allowed to be increased to with sample size re-estimation is 441. The dropout time D is generated from an exponential distribution with 5% of dropout per year. PFS and OS events will be followed for up to 2 and 4 years, respectively, from the enrollment of the last subject (if the event goal is not achieved). One-sided hypothesis testing is performed under the alternative hypothesis that PFS and OS are longer in the experimental group.

At the first interim analysis, only PFS will be tested. At the second interim analysis, PFS and OS will be tested separately. Due to the uncertainty in the treatment effect assumption at the initiation of the trial, it is desirable to have sample size re-estimation (SSR) based on the interim information to ensure power is achieved. The SSR approach provides some flexibility in case that the assumptions in the study design over-estimate treatment effect, and thus result in an under-powered study. If the interim analysis shows the conditional power of treatment effect is in a pre-specified range, promising zone, the sample size can be increased to achieve desired power. The sample size is increased using Gao et al. (2008) that allows sample size increasing proportional to the treatment effect at the interim. Cui et al. (1999) showed that type I error is inflated when sample size is

Table 6: Type I error for two end-points group sequential design

	PFS $\alpha = 0.005$	OS $\alpha = 0.02$
Mantel-Cox using normal approximation	0.00547	0.02076
Mantel-Cox using permutation test (5,000)	0.00473	0.01906
Peto-Peto using normal approximation	0.00419	0.01848
Peto-Peto using permutation test (5,000)	0.00505	0.01976

increased. The weighted test statistic was developed to preserve type I error. In this approach, test statistics of standard normal distribution is assumed in order to use the weighted test statistics to control type I error.

Type I error with the Mantel-Cox and Peto-Peto log-rank tests using a normal approximation and permutation test are evaluated. The results are shown in Table 6.

Consistent with previous observations from simple one end-point design, Table 6 shows that Peto-Peto log-rank test using normal approximation has smaller type I error than pre-determined α level for both PFS and OS, while Mantel-Cox log-rank test has inflated type I error. However, using an approximate permutation approach, the type I errors for both tests can be effectively controlled at the desired level.

6.2 Power

Although we have focused on controlling type I error, power is also an important operating characteristic. It is of interest that the approximate permutation test not only controls type I error, but also preserves power under the alternative hypothesis.

The parameters used for power analysis are the same as the simulation design in Section 6.1, except for OS and PFS time specifications. Under the alternative hypothesis, T_{os} in experimental and control groups is generated from exponential distribution with median survival time of 10 and 7 months respectively and hazard ratio of 0.7 (experimental vs. control). T_{pd} in experimental and control groups is generated from exponential distribution with median survival time of 34 and 14 months, respectively. As a result, median T_{pfs} is 7.7 months in experimental group and 4.7 months in control group with hazard ratio of 0.6 (experimental vs. control). Number of subjects and events are calculated such that power goals of PFS and OS are achieved. The power goal for PFS is 90% and the power goal for OS is 80%. Several interim analyses for OS and PFS have been planned in Table 7 based on this simulation design. PFS will be tested for futility at the first and second interim analysis. SSR is determined at second interim analysis and the trial will stop early if the third interim analysis shows enough evidence of treatment efficacy. 100,000 simulations are performed, and results are shown in Table 8.

From Table 8, we can see that, with either test based on an asymptotic normal distribution or approximate permutation test, the power goals for both PFS and OS are achieved with both Mantel-

Table 7: Approximate timing of interim analysis for PFS and OS (Section 6.2 Setting)

Analysis	Time (Months)*	PFS (Progression Free Survival)	OS (Overall Survival)
1	18.4	Interim for futility (50% events), use $265 \times 0.5 = 133$ events	——
2	25.4	Interim for futility and sample size re-estimation (80% events), use $265 \times 0.8 = 212$ events	Interim for futility and sample size re-estimation (60% events), use $304 \times 0.60 = 183$ events
3	29.9 (Enrollment completes)	Final (265 events)	Stopping early for efficacy (77% of events), use $304 \times 0.77 = 233$ events
4	38.3	——	Final (304 deaths)

*From the date of randomization of the first subject

Table 8: Power analysis for two end-points group sequential design

	PFS $\alpha = 0.005$ Power goal 90%	OS $\alpha = 0.02$ Power goal 80%
Mantel-Cox using normal approximation	0.91786	0.84322
Mantel-Cox using permutation test (5,000)	0.90888	0.83442
Peto-Peto using normal approximation	0.90138	0.82953
Peto-Peto using permutation test (5,000)	0.90885	0.83397

Table 9: Type I error for two end-points group sequential design with different survival time distribution

Weibull distribution (*Case 1*)	PFS $\alpha = 0.005$	OS $\alpha = 0.02$
Mantel-Cox using normal approximation	0.00620	0.02207
Mantel-Cox using permutation test (5,000)	0.00535	0.02016
Peto-Peto using normal approximation	0.00396	0.01833
Peto-Peto using permutation test (5,000)	0.00507	0.01998
Log-normal distribution (*Case 2*)		
Mantel-Cox using normal approximation	0.00610	0.02230
Mantel-Cox using permutation test (5,000)	0.00536	0.02040
Peto-Peto using normal approximation	0.00392	0.01835
Peto-Peto using permutation test (5,000)	0.00502	0.02001

Cox and Peto-Peto log-rank tests. Generally, using Peto-Peto log-rank test will lead to a loss of power, but approximate permutation tests for both Mantel-Cox and Peto-Peto log-rank tests achieve similar power.

6.3 Change of survival time distribution

In addition to the exponential distribution, Weibull and log-normal distribution are also assessed. The setup of simulations is the same as Section 6.1, except for survival time distribution and approximate timing of interim analysis. More specifically, we assessed the following two cases:

- *Case 1:* T_{os} follows a Weibull distribution with shape of 2, scale of 14. T_{pd} follows a Weibull distribution with shape of 2, scale of 24. Four interim analysis happen after 21.1, 27.6, 31.9 and 37.4 months.

- *Case 2:* T_{os} follows a log-normal distribution with $\mu = 2.5$, $\sigma = 0.5$. T_{pd} follows a log-normal distribution with $\mu = 3$, $\sigma = 0.5$. Four interim analysis happen after 22.5, 29.1, 33.4 and 38.7 months.

Type I error with the Mantel-Cox and Peto-Peto log-rank test using asymptotic normal distribution and permutation test is shown in Table 9.

Similar to the results from exponential distribution in Table 6, using asymptotic normal distribution, Mantel-Cox log-rank test has inflated type I error and Peto-Peto log-rank test has smaller type I error than predetermined α. With approximate permutation test, though inflation still present with Mental-Cox log-rank test, The type I errors were closer to pre-determined α level for both tests.

7 Discussions and Conclusions

We have identified several factors impacting type I error of the log-rank test, such as randomization allocation ratio between treatment groups. Although the most fundamental factor is the sample size, sample sizes required to control the type I error (i.e. those of several thousand patients) are often impossible to achieve in a clinical trial. Many trials are designed to demonstrate statistically significant differences with just a few hundred subjects, but we've demonstrated that the normal distribution is not a good approximation for the commonly used Mantel-Cox log-rank test statistic for this type of sample size, due to type I error inflation. Such inflation is not negligible with even moderately large sample sizes.

We investigate the performance of the Peto-Peto log-rank test as an alternative to control type I error with little loss of power. We also propose the approximate permutation test as an alternative approach. Simulations show that type I error can be controlled at pre-determined α levels with this approach. While more permutations provide a better approximation to the true distribution, it also increases computational burden. We recommend that 5,000 permutations are sufficient. The approach has been applied to a simple design with one end-point and no adaption and a complicated group sequential design with two endpoints and sample size re-estimation. It provides desirable operating characteristics of controlling type I error under the null hypothesis and provided desired power under the alternative hypothesis. Related tests can be found in R packages and . However, when assumptions are violated, such as informative censoring, neither test can overcome the associated bias. Although both approaches provide attractive operating characteristic, obtaining regulatory agreement is recommended when applying to clinical trials.

We mentioned at the beginning of Section 6, Cui et al. (1999) proposed a weighted test statistic at the final analysis to control type I error in case of sample size re-estimation. Weights are computed based on observed outcome and the resulting weighted test statistic is compared against a critical value determined by α spending function. Under this framework, boundaries are set based on standard normal distribution for the log-rank test statistic at interim and final analysis (Gordon Lan and Demets, 1983). When applying the approximate permutation test, the test statistic is not normally distributed. We have shown type I error is controlled through simulation. However, theoretical proof of the extension of Cui-Hung-Wang method is needed (Cui et al., 1999).

Acknowledgements

We are grateful for the computational resources for this paper provided by Center for Research Computing (www.crc.pitt.edu) and PittGrid (www.pittgrid.pitt.edu) from University of Pittsburgh.

References

Arboretti, R., Fontana, R., Pesarin, F., and Salmaso, L. (2018), "Nonparametric combination tests for comparing two survival curves with informative and non-informative censoring," *Statistical Methods in Medical Research*, 27, 3739–3769.

Berry, K. J., Johnston, J. E., and Mielke Jr, P. W. (2011), "Permutation methods," *Wiley Interdisciplinary Reviews: Computational Statistics*, 3, 527–542.

Brendel, M., Janssen, A., Mayer, C.-D., and Pauly, M. (2014), "Weighted Logrank Permutation Tests for Randomly Right Censored Life Science Data," *Scandinavian Journal of Statistics*, 41, 742–761.

Brown, M. (1984), "On the choice of variance for the log rank test," *Biometrika*, 71, 65–74.

Cui, L., Hung, H. M. J., and Wang, S.-J. (1999), "Modification of sample size in group sequential clinical trials," *Biometrics*, 55, 853–857.

Davison, A. C. and Hinkley, D. V. (1997), *Bootstrap methods and their application*, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.

Dwass, M. (1957), "Modified randomization tests for nonparametric hypotheses," *The Annals of Mathematical Statistics*, 28, 181–187.

Gao, P., Ware, J. H., and Mehta, C. (2008), "Sample size re-estimation for adaptive sequential design in clinical trials," *Journal of Biopharmaceutical Statistics*, 18, 1184–1196.

Gordon Lan, K. K. and Demets, D. L. (1983), "Discrete sequential boundaries for clinical trials," *Biometrika*, 70, 659–663.

Heimann, G. and Neuhaus, G. (1998), "Permutational distribution of the log-rank Statistic under random censorship with applications to carcinogenicity assays," *Biometrics*, 54, 168–184.

Heinze, G., Gnant, M., and Schemper, M. (2003), "Exact log-rank tests for unequal follow-up," *Biometrics*, 59, 1151–1157.

Heller, G. and Venkatraman, E. S. (1996), "Resampling procedures to compare two survival distributions in the presence of right-censored data," *Biometrics*, 52, 1204–1213.

Janssen, A. (1997), "Studentized permutation tests for non-i.i.d. hypotheses and the generalized Behrens-Fisher problem," *Statistics & Probability Letters*, 36, 9–21.

Janssen, A. and Pauls, T. (2003), "How do bootstrap and permutation tests work?" *Annals of statistics*, 768–806.

Kellerer, A. M. and Chmelevsky, D. (1983), "Small-sample properties of censored-data rank tests," *Biometrics*, 39, 675–682.

Klein, J. P. and Moeschberger, M. L. (2006), *Survival analysis: techniques for censored and truncated data*, Springer Science & Business Media.

Latta, R. B. (1981), "A monte carlo study of some two-sample rank tests with censored data," *Journal of the American Statistical Association*, 76, 713–719.

Mantel, N. (1988), "Evaluation of survival data and two new rank order statistics arising in its consideration," *Cancer chemotherapy reports*, 50, 163–170.

Neuhaus, G. (1988), "Asymptotically optimal rank tests for the two-sample problem with randomly censored data," *Communications in Statistics - Theory and Methods*, 17, 2037–2058.

— (1993), "Conditional rank tests for the two-sample problem under random censorship," *The Annals of Statistics*, 21, 1760–1779.

Peto, R. and Peto, J. (1972), "Asymptotically efficient rank invariant test procedures," *Journal of the Royal Statistical Society. Series A (General)*, 135, 185–207.

Schoenfield, D. (1981), "The asymptotic properties of nonparametric tests for comparing survival distributions," *Biometrika*, 68, 316–319.

Strawderman, R. L. (1997), "An asymptotic analysis of the logrank test," *Lifetime Data Analysis*, 3, 225.

Tang, Z. (2014), "Unequal randomization allocation and cox regression a simulation study," *Biometrics & Biostatistics International Journal*, 1.

Received: October 5, 2019

Accepted: January 11, 2020