Published in final edited form as:

Stat Methods Med Res. 2017 February ; 26(1): 75–87. doi:10.1177/0962280214539282.

Tests for Equivalence of Two Survival Functions: Alternative to The Tests Under Proportional Hazards

Elvis E. Martinez[1,*], **Debajyoti Sinha**[1], **Wenting Wang**[2], **Stuart R. Lipsitz**[3], and **Richard J. Chappell**[4]

[1]Department of Statistics, Florida State University, Tallahassee, U.S.A

[2]Department of Biostatistics, MD Anderson Cancer Center, University of Texas, Houston, U.S.A

[3]Harvard Medical School, Boston, U.S.A

[4]UW-Madison, Biostatistics & Medical Informatics, Madison, U.S.A

Abstract

For either the equivalence trial or the non inferiority trial with survivor outcomes from two treatment groups, the most popular testing procedure is the extension (e.g, Wellek[1]) of log-rank based test under proportional hazards model (PHM). We show that the actual type I error rate for the popular procedure of Wellek[1] is higher than the intended nominal rate when survival responses from two treatment arms satisfy the proportional odds survival model (POSM). When the true model is POSM, we show that the hypothesis of equivalence of two survival functions can be formulated as a statistical hypothesis involving only the survival odds-ratio parameter. We further show that our new equivalence test, formulation, and related procedures are applicable even in the presence of additional covariates beyond treatment arms, and the associated equivalence test procedures have correct type I error rates under the PHM as well as the POSM. These results show that use of our test will be a safer statistical practice for equivalence trials of survival responses than the commonly used log-rank based tests.

Keywords

Clinical importance; Cox's model; Critical region; Proportional Odds Survival Model

1 Introduction

Clinical trials for determining equivalence of a new treatment with a standard treatment of proven efficacy have become increasingly commonplace in recent years. With growing financial and ethical pressures (e.g., WMA Declaration of Helsinki[2]) to switch from an expensive and invasive standard treatment/procedure to a cheaper and less-invasive treatment, we can expect an increasingly higher number of equivalence trials to be conducted in future years. Wellek[1] using the proportional hazards model (PHM) of Cox[3]. The reason behind the popularity of this method for equivalence trial is given below. One

* elvism@stat.fsu.edu.

main challenge for developing a convenient hypothesis testing method for an equivalence trial is the formulation of the statistical hypothesis using only the parameter of the treatment effect. For a two-arm (placebo versus treatment) superiority trial under any semi-parametric model (e.g. the of Cox[3]), it is straightforward to make a Statistical methods used for equivalence trial for survival response are often based on methods of PHM statistical/ mathematical formulation of the alternative hypothesis H_a (clinically important difference) of scientific interest. Any effect of the treatment arm, denoted by non-zero regression parameter ($\eta \neq 1$) implies some difference $S_1(t) \neq S_0(t)$ in two survival curves at least at one time point t, and the converse is also true. For example, when two treatment arms follow proportional hazards model (PHM of Cox[3]) with hazard ratio η, the alternative hypothesis H_a: $S_1(t) \neq S_0(t)$ for some t, implies $H_a^*: \eta \neq 1$ and vice versa. However, for an equivalence trial, when the alternative H_a is $|S_1(t) - S_0(t)|$ being within the prespecified range of equivalence for every time-point t (to be explained later), it is not straightforward to express this H_a as a statistical hypothesis H_a^* involving only the regression parameter η (which is free of time t). For example, in Cox's PHM, it is not obvious that $|S_1(t) - S_0(t)|$ less than a small known constant for all t does imply that η is within a known interval. Wellek[1] paved the way for a convenient log-rank based equivalence test by deriving this result for the PHM, and only for the case of no covariates beyond treatment arms. Our result, an extension of the result of Wellek[1] to the case of POSM, allows us to formulate an equivalence test for the POSM based Wellek[4] (section 6.7) for a thorough review of the justifications behind formulating statistical hypothesis of equivalence based on the treatment effect parameter.

Due to Wellek[1] results, the existing literature on equivalence trials for survival responses is dominated by the log-rank test based on the assumption of on a rejection region which only involves the estimate and the corresponding standard error of the treatment effect parameter. Please see a PHM for the two treatment arms, without any consideration for alternative semi-parametric models and the presence of other covariates. The non parametric procedures of Com-nogue *et al.*[5] and others often require much higher sample sizes than tests based on semi-parametric models. In practice, often the hazard functions of two treatment arms are not proportional over time and there may be other covariates in addition to treatment arms. We show that a log-rank based test of equivalence has a higher than intended type I error rate when treatment arms do not follow the PHM. This practical need to consider new equivalence tests based on other semi-parametric models. For example, the ratio of two hazards may converge towards one over time when the initial benefit of one treatment arm over the other treatment arm diminishes over time. In this situation, the proportional odds survival model (POSM) of Bennett[6] will be more appropriate than a PHM. In this paper, we also show that a POSM based equivalence test has correct type I error even when the true model is either POSM or PHM. This shows that the POSM based equivalence test is a safer option in practice compared to the log-rank based test, especially when the underlying

We place high emphasis on controlling the type I error rate for an equivalence trial because, unlike for a superiority trial, an effective standard treatment already exists for an equivalence trial. Wrongly accepting the alternative H_a of equivalence can potentially replace an effective standard treatment with an ineffective treatment in the market. Whereas, even if we wrongly emphasizes the modeling assumption is under suspicion. accept the null of non

equivalence, i.e, do not accept the new treatment as equivalent, we will still have the effective standard treatment available in the market. In this case, wrongly rejecting the null is a more serious mistake than wrongly accepting the null. However, we first deal with a major impediment for developing an equivalence test for a POSM. The clinicians and other non statisticians have understandable difficulty in defining the clinically important difference between the two treatment arms in terms of the ratio of two survival odds. On the contrary, most clinical experts and researchers are comparatively more at ease the clinical equivalence of two treatment arms in terms of a clinically important difference between two survival functions. The development of equivalence trial methodology for POSM depends on whether the alternative hypothesis of the equivalence of two survival curves (or two hazard curves) can be properly expressed as an alternative hypothesis in the regression parameter of the POSM.

In Section 2, we first derive the formulation of the alternative statistical hypothesis, H_a^*, that only uses the odds ratio of the POSM, such that H_a^* also corresponds to the scientific (clinical) hypothesis related to the "equivalence" of the survival functions of two treatment arms. In section 3, we describe the statistical methods including rejection regions for two-sample and one-sample equivalence studies under POSM. In section 4, we show that even in the presence of additional covariates, testing equivalence of the survival functions for two treatment arms is the same as statistically testing the survival odds ratio parameter to be within a small interval. This result allows us to develop to express the statistical test of equivalence of two treatments under POSM, even in the presence of additional covariates. In section 5, we study the relationship between sample size and intended type I error rates with tests based on Cox's model and our new POSM based tests. Our theoretical and simulation studies show that when the POSM assumption is true for the trial in question, log-rank based equivalence test of Wellek[1] tends to reject the correct null hypothesis more often than the desired level of significance. On the contrary, our POSM based equivalence tests achieve desired type I error rates and power when the true model is either POSM or Cox's model.

2 Formulation of Hypothesis under POSM

For the time being, we consider no covariate other than treatment arm. We later extend our methods to include other covariates. The POSM of Bennett[6] assumes

$$\frac{1-S_1(t)}{S_1(t)}=\theta\left[\frac{1-S_0(t)}{S_0(t)}\right], \quad (2.1)$$

for all time points $t > 0$, where θ is the time-constant survival odds ratio between new treatment and standard treatment. With corresponding survival functions $S_1(t)$ and $S_0(t)$ respectively. For example, one may consider two treatments are clinically equivalent if $|S_1(t) - S_0(t)|$, the difference between two survival functions, is smaller than a predetermined equivalence level δ over time. Thus two treatment arms are equivalent only when $|S_1(t) - S_0(t)| < \delta$ for all t. Here, the additional quantity $\delta > 0$ indicates the maximum clinical difference allowed between the standard therapy and a therapeutically equivalent

experimental therapy. The value of δ is usually determined by clinical experts and regulatory agencies involved in determining the practical definition of the equivalence of two treatments under consideration. However, in order to implement a statistical test for the equivalence of two treatments under POSM of (2.1), the alternative statistical hypothesis H_a^* must be based on a range (interval) of θ, where the interval depends on the practical (clinical) meaning of the equivalence of two survival curves $S_1(t)$ and $S_0(t)$. Furthermore, it is difficult for clinicians and non statisticians to express the therapeutic equivalence in terms of a prespecified range of θ, because θ is a ratio of odds, unlike difference in probabilities of any observable event under two treatment arms. To facilitate the formulation of a statistical hypothesis testing procedure for evaluating the clinical (scientific) alternative hypothesis H_a: $|S_1(t) - S_0(t)| < \delta$ for all t, under POSM of (2.1) we develop the following theorem.

Theorem 1

Under POSM of (2.1) with continuous $S_0(t)$, testing H_a: $|S_1(t) - S_0(t)| < \delta$ for all $t > 0$, is the same as testing $H_a^*: (1+\varepsilon)^{-1} < \theta < 1+\varepsilon$, where $\varepsilon = (4\delta)/(1 - \delta)^2$ is a known function of δ.

Theorem 1 (proof in the Appendix (A-1)) shows that under the POSM of (2.1), if the clinicians and practitioners can specify the maximum allowable difference δ between two survival functions $S_1(t)$ and $S_0(t)$ of two equivalent treatment arms, we can derive the corresponding statistical alternative hypothesis H_a^* based on the time-constant survival odds ratio θ. This H_a^* can now be tested using statistical hypothesis testing tools.

Many authors including Rothman *et al.*[7] advocated testing the equivalence of two treatments using the hazard ratio because the hazard ratio of Cox's model does not depend on the baseline population. The hazard ratio is also the popular parameter for comparing treatments in efficacy trials (at least in the field of oncology). One may specify the alternative (scientific) hypothesis H_a of equivalence of the two treatment arms via H_a: $|h_1(t)/h_0(t)| < \rho$ for all time points $t > 0$, where $h_1(t)$ and $h_0(t)$ are hazard functions for new and standard treatments respectively. Similar to δ for Theorem 1, the maximum allowable hazard ratio $\rho > 1$ for two clinically equivalent treatments is determined from a clinical perspective. To expedite the equivalence trial under POSM of (2.1) for H_a based on hazards ratio, we have the following theorem (proof is again in the Appendix (A-2)).

Theorem 2

Under the POSM assumption of (2.1), the alternative hypothesis of interest H_a: $|h_1(t)/h_0(t)| < \rho$ for all t, is the same as testing $H_a^*: \rho^{-1} < \theta < \rho$.

We note that the H_a^* here is identical to the H_a^* of Result 1 with $(1 + \varepsilon)$ replaced by ρ. This indicates that for POSM of (2.1), the formulation of the statistical hypothesis H_a^* is the same while testing the equivalence of two treatment arms based on either the maximum hazards ratio over time or the maximum difference of the survival functions over time. Both of these alternative hypothesis can be reduced to testing the statistical hypothesis H_a^* involving only time constant parameter θ in (2.1). In the next section, we present the statistical tests and corresponding critical regions for this hypothesis H_a^* for two cases – the two-sample case

when the baseline survival function $S_0(t)$ of standard treatment is unknown and the one-sample case when $S_0(t)$ is known from historical data.

3 Implementation of Equivalence Tests

First we discuss the statistical tests for the equivalence of two treatment arms under the POSM of (2.1) when n patients are randomized to two treatment arms with $z_i = 1$ when patient i receives the new treatment, and $z_i = 0$ when she/he receives the standard treatment. We denote the observed right-censored data as $(\underset{\sim}{Y}, \underset{\sim}{\delta}, \underset{\sim}{z})$, where $\underset{\sim}{Y} = (Y_1, \ldots, Y_n)$ and observed censoring indicators $\underset{\sim}{\delta} = (\delta_1, \ldots, \delta_n)$ where Y_i is the observed survival when $\delta_i = 1$ and Y_i is the right-censoring time when $\delta_i = 0$. Survival time T_i is at risk of non informative random right censoring. In practice, the decision about therapeutic equivalence of two treatment arms will be based on testing H_0: $|S_1(t) - S_0(t)| \geq \delta$ for some time-point t, versus H_a: $|S_1(t) - S_0(t)| < \delta$ for all $t > 0$. From Theorem 1 and (2.2), we know that testing this hypothesis is equivalent to testing

$$H_0^*: |\beta| \geq \log(1+\varepsilon) \text{ versus } H_a^*: |\beta| < \log(1+\varepsilon), \quad (3.1)$$

where $\varepsilon = (4\delta)/(1-\delta)^2$ and $\beta = \log(\theta)$ in (2.1). Due to the formulation of this statistical equivalence test based solely on parameter β of POSM, we can use the test statistic of a superiority test under POSM as the test statistic for testing (3.1). However, the new test for (3.1) has a different rejection region.

One option is to use the semi-parametric Maximum likelihood estimator (SPMLE) $(\hat{\beta}, \hat{B})$ of Murphy *et al.,*[8] obtained via maximizing the following semi-parametric likelihood

$$L(\beta, B_0 | \underset{\sim}{Y}, \underset{\sim}{\delta}) \propto \prod_{i=1}^{n} \left(\frac{exp(z_i\beta)}{B_0(Y_i) + exp(z_i\beta)} \right) \left(\frac{\Delta B_0(Y_i)}{B_0(Y_i-) + exp(z_i\beta)} \right)^{\delta_i},$$

where the baseline odds function $B_0(t) = S_0(t)/\{1-S_0(t)\}$ is a non decreasing, right continuous function and with jumps $\Delta B_0(t) = B_0(t) - B_0(t-)$ at the observed failure times. The rejection region of the large sample based asymptotically most powerful test for (3.1) is given as:

$$\left\{ |\hat{\beta}| \sqrt{\hat{I}_\beta} < C_\alpha \left(\sqrt{\hat{I}_\beta} \log(1+\varepsilon) \right) \right\}, \quad (3.2)$$

where the $C_\alpha(\psi)$ is the square root of the αth quantile of a χ^2 distribution with $df = 1$ and non centrality parameter ψ^2. Numerical differentiation of the profile likelihood $\text{prlik}_n = \log\{L(\beta, \hat{B}_0 | \mathbf{Y}, \boldsymbol{\beta})\}$ is used to obtain

$$\hat{I}_\beta \approx -\frac{1}{nh^2}\{\text{prlik}_n(\hat{\beta}+h)-2\text{prlik}_n(\hat{\beta})+\text{prlik}_n(\hat{\beta}-h)\},$$

for some small enough h given in Murphy *et al.*[8].

An alternative semi-parametric approach for testing (3.1) is to use the test statistic of Chen *et al.,*[9] based on the estimator $\tilde{\beta}$ obtained via iteratively solving a set of estimating equations. The iterative steps are outlined in Appendix (A-3). Using the test statistic of Chen *et al.,*[9] we can similarly derive the rejection region for testing (3.1) as

$$\left\{\frac{|\tilde{\beta}|}{\nu(\tilde{\beta})}<C_\alpha\left(\frac{\log(1+\varepsilon)}{\nu(\tilde{\beta})}\right)\right\}, \quad (3.3)$$

where $C_\alpha(\psi)$ is the square-root of the α-quantile of χ^2 with df=1 and non centrality parameter ψ^2. By taking this approach, we can avoid high dimensional numerical maximization and the estimator $\nu^2(\tilde{\beta})$ of the asymptotic variance of $\tilde{\beta}$ has a closed-form expression. We omit the closed form expression of $\nu(\tilde{\beta})$ (given in Chen *et al.*[9]) for the sake of brevity. Although the estimator $\tilde{\beta}$ is not the most efficient estimator, the efficiency loss is typically small.

In many equivalence trials, particularly in oncology, for all practical purposes, we may know the baseline survival $S_0(t)$ of the standard treatment. In particular, there often exists a considerable amount of historical data on the survival function $S_0(t)$ of the standard treatment because its efficacy has been already studied. In this situation, every patient with observed survival data Y_i and censoring indicator δ_i for $i = 1, \ldots, n$ receives the new treatment. 11 We note that the logic and the result of Theorem 1 still apply here and the hypothesis of equivalence of two treatment is again reduced to the hypothesis of (3.1). Since $S_0(t)$ is known, we can find the MLE ($\hat{\beta}$) of β by solving the score equation:

$$n-\sum_{i=1}^{n}(1+\delta_i)\frac{exp(\beta)}{exp\{-B(y_i)\}+exp(\beta)}=0,$$

where $B(t) = S_0(t)/\{1 - S_0(t)\}$ is known. Using the usual asymptotic theory, the large-sample rejection region is $\frac{|\hat{\beta}|}{\nu(\hat{\beta})}<C_\alpha\left(\frac{\log(1+\varepsilon)}{\nu(\hat{\beta})}\right)$, where $C_\alpha(\psi)$ is the same as in (3.3) and the estimated variance $\nu^2(\hat{\beta})$ has the closed form expression

$$\nu^2(\hat{\beta})=\sum_{i=1}^{n}(1+\delta_i)\frac{B_0(y_i)exp(\hat{\beta})}{n\{B_0(y_i)+exp(\hat{\beta})\}^2}.$$

The computer codes for computing the test statistics of (3.2) and corresponding critical regions of (3.3) are available from the authors upon request. The authors also have codes for the competing test statistics and critical regions of Wellek[1].

4 Extension to include other covariates

We now extend our previously described procedure of equivalence tests to accommodate even other covariates $\boldsymbol{x}_j$, in addition to treatment arm indicator z_j. Even though it is very much conceivable to have additional covariates in practice, we have not yet come across any previous research on equivalence 12 tests to accommodate additional covariates. We assume that the underlying model is a natural extension of the POSM of (2.1) with

$$\frac{1-S_1(t|\boldsymbol{x})}{S_1(t|\boldsymbol{x})}=\theta\left[\frac{1-S_0(t|\boldsymbol{x})}{S_0(t|\boldsymbol{x})}\right]=\theta e^{\boldsymbol{\gamma x}}\left[\frac{1-S_0(t)}{S_0(t)}\right], \quad (4.1)$$

where $\boldsymbol{\gamma}$ is the regression parameter of $\boldsymbol{x}$ and θ is again the treatment effect of interest. The clinical hypothesis of interest, H_a: $|S_1(t|\boldsymbol{x}) - S_0(t|\boldsymbol{x})| < \delta$ for all covariates $\boldsymbol{x}$ and for all $t > 0$, is equivalent to testing the statistical hypothesis $H_a^*:(1+\varepsilon)^{-1}<\theta<(1+\varepsilon)$, where $\varepsilon = (4\delta)/(1 - \delta)^2$ (proof omitted). It is important to note that H_a^* does not depend on either $\boldsymbol{\gamma}$ or $\boldsymbol{x}$. This result shows that for survival response with the POSM assumption, the hypothesis of equivalence of two patients with the same covariate $\boldsymbol{x}$ but from different treatment arms is the same as testing the statistical hypothesis H_a^*. This result allows us to extend the formulation of the statistical hypothesis of equivalence in Theorem 1 to the equivalence studies under POSM with additional covariates $\boldsymbol{x}$. However, the test statistic and corresponding critical region are now different from those used for equivalence tests with no covariates. The new test statistic, its corresponding critical region, and associated computational steps are given in the Appendix (A-4).

5 Error Rates of Tests

Since the properties of our equivalence testing procedures do not depend on additional covariates $\boldsymbol{x}$, for the sake of simplicity, we do not include covariate $\boldsymbol{x}$ for our theoretical and simulation studies to compare the error rates of competing procedures. In this section, we first theoretically show inflation of type I error rate of the PHM based test when true model is POSM. After that, we also perform simulation studies to study the finite sample properties (type I error and power) of both the POSM-based tests and the log-rank based tests under correctly and incorrectly specified models.

In practice, the most frequently used semi-parametric procedure for testing the equivalence (e.g. Wellek[1]) is via a log-rank based statistic under the assumption of the proportional hazards model (PHM) of Cox[3]:

$$h_1(t)/h_0(t)=exp(\eta), \quad (5.1)$$

where $h_0(t)$ is the baseline hazard and $exp(\eta)$ is the hazards ratio of the two treatment arms under the PHM. In spite of substantial literature on the robustness of a log-rank statistic based on the PHM of (5.1) for superiority tests, there is not much research studying the effect of wrongly using a log-rank based test statistics for an equivalence hypothesis when the true underlying model is not of (5.1). We examine the type I error rate for wrongly using a log-rank based equivalence test when the true underlying model is the POSM of (2.1) with true value of β as $\beta_0 = 2\log\{(1+\delta)/(1-\delta)\}$. It means 14 that two treatment arms following the POSM of (2.1) have the maximum difference of δ between their survival curves. If we wrongly use a log-rank based equivalence test with the same δ, we actually use a test based on the partial likelihood estimate $\hat{\eta}$ of Cox[3]. In this case, the asymptotic density of $\hat{\eta}$ is not centered around true parameter value β_0 of model (2.1). Instead, Lin and Wei[10] showed that $n^{1/2}(\hat{\eta}-\eta^*)$ follows an asymptotic normal distribution with mean 0 and variance $v^2(\eta)$, where η^* is the unique solution of the equation

$$n_1-\int_0^{+\infty}\frac{n_1e^{\eta}S_0(t)}{n_1e^{\eta}S_0(t)+n_0S_1(t)}dt=0, \quad (5.2)$$

and where n_0 and n_1 are the sample sizes for the standard treatment and new treatment respectively. Here, $v(\eta)$ is the estimated standard error of $\hat{\eta}$ obtained from Cox[3]. When the sample sizes n_0 and n_1 in the two treatment arms increase to $+\infty$, we can show that the center of the asymptotic distribution of η is $|\eta| < \log(1+\varepsilon_h)$, where ε_h satisfies $(1+\varepsilon_h)^{-1/\varepsilon_h}-(1+\varepsilon_h)^{-(1+\varepsilon_h)/\varepsilon_h} = \delta$ (the proof is in the Appendix (A-5)). Since the rejection region for the logrank based test is

$$\left\{\frac{|\eta|}{v(\eta)}<C_\alpha\left(\frac{\log(1+\varepsilon_h)}{v(\eta)}\right)\right\},$$

the necessary condition for controlling the type I error rate within 0.05 for large sample size is $|\eta^*| = \log(1+\varepsilon_h)$. Under the null hypothesis H_0, as sample sizes become sufficiently large and $|\eta^*|$ goes below $\log(1+\varepsilon_h)$, the type I error rate for a log-rank based test becomes greater than 0.05, the intended type I error rate of the test. Below, we also show, via simulation studies, the approximate levels of inflation of the type I error rate for finite sample sizes if we wrongly use a log-rank based test when the true model is POSM of (2.1) with true regression parameter β_0.

Our simulation studies with underlying POSM use $S_0(t) = \Phi(2-\log(t))$, a log normal baseline survival function with mean = 2 and variance = 1, and The test-statistics for the log-rank and POSM based tests were calculated in Matlab. We take the maximum allowable difference in survival curves between two equivalent treatments as $\delta = 0.15$, the same used by Wellek[1]. Using Theorem 1, we get the corresponding $\varepsilon = 0.8304$, the cut-off for the equivalence test based on POSM. Each entry gives the fraction of times out of 1, 000 replications of simulated data sets for which the test statistic falls in the critical region of (3.2) with $\delta = 0.15$ (that is $\varepsilon = 0.8304$). The columns for $m = max|S_1(t) - S_0(t)| = 0$ and 0.10

represent the approximate powers of the tests. The rest of the columns represent the type I error rates (sizes) of the tests at different $m \geq 0.15$. Table 1 shows the approximate powers and sizes using the POSM test, it appears to be below the nominal significance level of 0.05.

Table 2 summarizes the approximate powers and sizes for (wrongly) using the log-rank test proposed by Com-Nougue *et al.*[5] and Wellek[1] for equivalence using the same 1, 000 replicate data sets simulated from the POSM models. We use the rejection region of Wellek[1], with intended test size 0.05 and the maximum difference in survival curves $\delta = 0.15$ as the margin of equivalence an exponential with mean 50 random censoring variable. (same as Table 1). Each entry gives the fraction of replications for which the test statistic falls in the critical region of Wellek[1] for $\delta = 0.15$. The simulation results show that the type I error rates at the boundary of the null H_0 of the log-rank based tests are greater than 0.05 when the true model is POSM. The difference between the actual (estimated) size type I error rate and the intended probability of type I error (5%) increases as the sample size increases. This indicates that when we wrongly use a log-rank based test, the probability of accepting the alternative that the two treatments are equivalent even when they are actually different from each other (null is true) is higher than the intended level of significance of the test.

Next, we perform a simulation study using data sets generated from the proportional hazards model (PHM), where $S_0(t) = \Phi(2-\log(t))$ and random censoring following an exponential with mean 50. We again compare the type I error rates as well as the power of the log-rank based test with those of our POSM based tests. The rejection regions for both tests are determined using the equivalence margin $\delta = 0.15$ and an intended level of significance of 0.05. The values in Table 3 are the results using the log-rank based test under PHM. Table 4 values represent the POSM based test when the simulation model is PHM. Although we have only limited amount of loss of power for wrongly assuming the POSM compared to the powers of the log-rank based test, the type I error rates (sizes) of POSM based test remain below and close to the intended 5% level. This shows that the test based on a POSM assumption is a more conservative and robust approach, when compared to the log-rank based test, even when the true underlying model has proportional hazards.

For the sake of brevity, we skip the results of the simulation study of the one sample case comparing the PHM based test and the POSM based test. Similar to the two-sample case, the size of our POSM based one-sample test has type I error lower than intended significance level test even when the sample size is small. The power of the one-sample test is almost double compared to the corresponding power of the two-sample test, indicating that we need a smaller number of patients compared to the two sample case when $S_0(t)$ is known.

6 Data Example & Conclusion

The main goal of the pediatric oncology trial of Nam *et al.*[11] was to evaluate whether a seven months long maintenance treatment (the standard) is equivalent to a shorter (less toxic) four months long treatment (new treatment) for non-Hodgkin's malignant type B lymphoma. It is necessary to accept a "small" decrease in survival rate as a trade-off for the better tolerability and less toxicity of a new shorter treatment. For this study, the investigators decided that $\delta =$

0.09 would be the “threshold of equivalence region” from Nam *et al.*[11] - the maximum difference between the survival rates of two equivalent treatment arms. Using this $\delta = 0.09$, the p-value of the log-rank based equivalence test was 0.024 based on the observed data with 11 and nine failures out of sample sizes of 84 and 82 respectively from standard long and new shorter treatment regimens. This p-value ($0.024 < 0.05$) is a highly significant evidence in favor of the alternative hypothesis of equivalence of two treatments. We cannot re-analyze the clinical trial data because the original data is proprietary. Instead, we would like to demonstrate using a simulation study the comparison between a log-rank based test and a POSM based test when we follow the design and the censoring mechanism similar to this trial and the true difference between treatment arms is at the margin of equivalence ($m = max|S_1(t) - S_0(t)| = 0.09$). We simulate 1000 data sets with 11 out of $n_1 = 84$ and nine failures out of $n_2 = 82$ for two treatment arms following POSM. We use the censoring scheme and monitoring length (18 months) similar to Nam *et al.*[11]. The proportion of log-rank based test-statistics with p-values more extreme than 0.024 is 0.038. When using the POSM test on the same set of data, we find the corresponding proportion to be 0.012. This shows that when the true model is POSM, the probability that a log-rank test will have a p-value less than 0.024 (same or more significant than the p-value obtained by Nam *et al.*[11]) is almost three times more than the probability of having such a significance level with the POSM based test. This further demonstrates that a log-rank based test has a high probability to give a highly-significant (very low) p-value, even though we would not like to reject the null hypothesis H_0 when the actual $\delta = 0.09$.

Unlike Cox’s model where one hazard function dominates the other over time, a POSM can allow two hazard functions to merge over time. This may be a possible explanation for POSM based equivalence test being more conservative than the log-rank based test. Li *et al.*[12] and Betensky *et al.*[13] (among others) have argued that even for superiority trials, the efficiency and validity of a log-rank based test likely is questionable for solid tumor oncology studies in which there is heterogeneity of tumors due to existence of various unidentified genetic subtypes. For these studies, the hazard functions from two treatment arms may merge over time. A POSM assumption with POSM based equivalence tests would be a wise choice in this situation.

In this paper, we have presented the test statistics, critical region and robustness and other related properties of POSM based tests restricted only to right-censored survival data. We would like point out that the statements of Theorem 1 and Theorem 2 are valid irrespective of the censoring mechanism. Arguably, for any type of censoring, it is possible to develop an equivalence test based on POSM if one can determine the appropriate statistic (preferably based on the SPML estimate of β of POSM from interval-censored data) and the corresponding critical region. However, equivalence trials with different types of censoring such as interval-censoring are beyond the scope of this paper. These are important topics of future research.

Acknowledgments

We would like to thank the editor, and two referees for their valuable suggestions that led to the great improvement of our article. Authors are grateful for the support provided by NIH grant R01CA69222 for this research.

References

1. Wellek S. A log-rank test for equivalence of two survivor functions. Biometrics. 1993; 49:877–881. [PubMed: 8241376]
2. World Medical Association Declaration of Helsinki. Ethical principals for medical research involving human subjects. JAMA. 2000; 284:3043–3045. [PubMed: 11122593]
3. Cox DR. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society, Series B. 1972; 30:248–275.
4. Wellek, S. Testing statistical hypothesis of equivalence and non inferiority. 2. Chapman & Hall; Florida: 2010.
5. Com-Nougue C, Rodary C, Patte C. How to establish equivalence when data are censored: A randomized trial of treatments for B non Hodgkin lymphoma. Statistics in Medicine. 1993; 12:1353–1364. [PubMed: 8210831]
6. Bennett S. Analysis of Survival Data by the Proportional Odds Model. Statistics in Medicine. 1983; 2:273–277. [PubMed: 6648142]
7. Rothmann M, Li N, Chen G, Chi GYH, Temple R, Tsou H-H. Design and analysis of non inferiority mortality trials in oncology. Statistics in Medicine. 2003; 22:239–264. [PubMed: 12520560]
8. Murphy SA, Rossini AJ, Van Der Vaart AW. Maximum likelihood estimation in the proportional odds model. Journal of the American Statistical Association. 1997; 92:968–976.
9. Chen K, Jin Z, Ying Z. Semi-parametric analysis of transformation models with censored data. Biometrika. 2002; 89:659–668.
10. Lin DY, Wei LJ. The robust inference for the Cox Proportional Hazards Model. Journal of the American Statistical Association. 1989; 89:659–668.
11. Nam J, Kim J, Seungyeoun L. Equivalence of two treatments and sample size determination under exponential survival model with censoring. Computational Statistics and Data Analysis. 2005; 49:217–226.
12. Li Y, Adelstein DJ, Adams GL, Wagner H, Kish JA, Ensley JF, Schuller DE, Forastiere AA. An intergroup phase III comparison of standard radiation therapy and two schedules of concurrent chemoradiotherapy in patients with unresectable squamous cell head and neck cancer. Journal of Clinical Oncology. 2003; 21:92–98. [PubMed: 12506176]
13. Betensky RA, Louis DN, Cairncross JG. Influence of unrecognized molecular heterogeneity on randomized clinical trials. Journal of Clinical Oncology. 2003; 20:2495–2499.

APPENDIX

(A-1) Proof of Theorem 1

For the POSM of (2.1), $|S_0(t) - S_1(t)| < \delta$ for all t, $\Leftrightarrow$ $|\{1 + B_0(t)\}^{-1} - \{1 + B_0(t)\theta\}^{-1}| < \delta$ for all t, where $B_0(t) = \{1 - S_0(t)\}/S_0(t)$. When $B_0(t)$ is continuous, using standard calculus, we can show that

$$\max_{t>0} = |\{1+B_0(t)\}^{-1} - \{1+B_0(t)\theta\}^{-1}| = \left| \frac{\theta^{\frac{1}{2}} - \theta^{-\frac{1}{2}}}{(1+\theta^{\frac{1}{2}})(1+\theta^{-\frac{1}{2}})} \right| = M(\theta). \quad (1)$$

We can see that $M(\theta)$ is a decreasing (increasing) function when $\theta \in (0, 1)$ (when $\theta \in (1,+\infty)$). Therefore, the condition $|S_0(t) - S_1(t)| < \delta$ for all t, is equivalent to $M(\theta) < \delta \Leftrightarrow (1 + \varepsilon)^{-1} < \theta < 1 + \varepsilon$, where ε should satisfy $\frac{(1+\varepsilon)^{\frac{1}{2}} - (1+\varepsilon)^{-\frac{1}{2}}}{\{1+(1+\varepsilon)^{\frac{1}{2}}\}\{1+(1+\varepsilon)^{-\frac{1}{2}}\}} = \delta \Rightarrow \varepsilon = (4\delta)/(1-\delta)^2$.

(A-2) Proof of Theorem 2

Let $h_0(t)$ and $h_1(t)$ be the hazard function for $S_0(t)$ and $S_1(t)$, respectively. POSM of (2.1) implies $h_1(t)/h_0(t) = [1+(\theta - 1)S_0(t)]^{-1}$, which is an increasing (a decreasing) function of t converging to 1 when $\theta > 1$ (when $\theta < 1$). This implies that, the maximum of $|\log\{h_1(t)/h_0(t)\}|$ for all $t > 0$, is equal to $|\log\{1 + (\theta - 1)S_0(0)\}| = |\log(\theta)|$ Therefore, testing $|\log\{h_1(t)/h_0(t)\}| < \log(\rho)$ for all $t > 0$, is equivalent to testing $|\log(\theta)| < \log(\rho) \Leftrightarrow \rho^{-1} < \theta < \rho$.

(A-3)

Chen *et al.*[9] estimates $(\tilde{\beta}, \tilde{B})$ of (β,B) can be obtained by solving a set of estimating equations, where $B(t)$ is a non decreasing and non negative function with jumps only at observed failure time $t_{(1)} < \ldots < t_{(k)}$.

To compute the estimators, one can use the following iterative algorithm:

Step 0. Choose an initial value $\beta^{(0)}$ for β.

Step 1. Obtain $B^{(0)}(t_{(1)})$ by solving

$$\sum_{i=1}^{n} \frac{Y_i(t_{(1)})\exp(z_i\beta^{(0)})}{\exp(z_i\beta^{(0)})+\exp\{-B(t_{(1)})\}}=1,$$

where $Y_i(t)$ is the indicator that subject i is under observation at time t. Then obtain $B^{(0)}(t_{(k)})$, for $k = 2, \ldots, K$, sequentially by solving:

$$\sum_{i=1}^{n} \frac{Y_i(t_{(k)})\exp(z_i\beta^{(0)})}{\exp(z_i\beta^{(0)})+\exp\{-B(t_{(k)})\}}=1+\sum_{i=1}^{n} \frac{Y_i(t_{(k)})\exp(z_i\beta^{(0)})}{\exp(z_\beta^{(0)})+\exp\{-B(t_{(k-)})\}}.$$

Step 2. Obtain new estimate $\beta(1)$ of β by solving

$$U(\beta, B) \equiv \sum_{i=1}^{n} z_i[\delta_i - \Lambda\{B(t)+z_i\beta\}]=0,$$
$$\sum_{i=1}^{n} [\delta_i - \Lambda\{B(t)+z_i\beta\}]=0,$$

where $\Lambda(u) = \{1 + \exp(-u)\}^{-1}$.

Step 3. Repeat Step 1 and Step 2 until the prescribed convergence criteria are met.

(A-4) Estimating Equations with covariates

The estimates $(\tilde{\beta}, \tilde{\gamma}, \tilde{B})$ of (β, γ, B) of (4.1) (based on Chen *et al.*[9]) obtained via solving the estimating equations:

$$U(\beta,\boldsymbol{\gamma},B)\equiv\sum_{i=1}^{n}\int_0^{\infty}[z_i;\boldsymbol{x}_i][dN_i(t)-\tilde{Y}_i(t)d\Lambda\{B(t)+z_i\beta+\boldsymbol{x}_i'\boldsymbol{\gamma}\}]=0,$$
$$\sum_{i=1}^{n}[dN_i(t)-\tilde{Y}_i(t)d\Lambda\{B(t)+z_i\beta+\boldsymbol{x}_i'\boldsymbol{\gamma}\}]=0. \quad (2)$$

The main iterative steps have been omitted for the sake of brevity.

(A-5) Proof for $|\eta^*| < \log(1 + \varepsilon h)$ when $n0, n1 \to +\infty$

For brevity, we only show the proof when $n_0 = n_1$. The equation of (5.2) converges to

$$1-\int_0^{+\infty}\frac{e^{\eta^*}S_0(t)}{S_1(t)+e^{\eta^*}S_0(t)}\left\{-S_1'(t)-S_0'(t)\right\}dt=0, \quad (3)$$

where $S_1(t) = S(t|z = 1)$, $S_0(t) = S(t|z = 0)$ follow POSM of (2.1), and $S_1'(t)$ and $S_0'(t)$ are the derivatives of $S_1(t)$ and $S_0(t)$ respectively.

Using $b_0(t) = dB_0(t)/dt$, in (4.2), we get $S_0(t) = \{1+e^{\beta_0}B_0(t)\}^{-1}$, $S_1(t) = \{1 + B_0(t)\}^{-1}$ and $-S_1'(t)=\frac{e^{\beta_0}b_0(t)}{\{1+e^{\beta_0}B_0(t)\}^2}$, $-S_0'(t)=\frac{b_0(t)}{\{1+B_0(t)\}^2}$. Using these $S_j(t)$ and $-S_j'(t)$ for $j = 1, 2$ in (3), we can get the equation for η^* as

$$1-\int_0^{+\infty}\frac{\{1+B_0(t)\}e^{\eta^*}}{\{1+B_0(t)e^{\beta_0}\}+\{1+B_0(t)\}e^{\eta^*}}\times\left\{\frac{b_0(t)}{\{1+B_0(t)\}^2}+\frac{b_0(t)e^{\beta_0}}{\{1+e^{\beta_0}B_0(t)\}^2}\right\}dt=0. \quad (4)$$

With further Calculus manipulation, we show η^* to be the unique solution of

$$U(\eta)=e^{\eta}\log\left\{\frac{1+e^{\eta}}{e^{\beta_0}+e^{\eta}}\right\}-e^{\beta_0-\eta}\log\left\{\frac{e^{\beta_0}+e^{\eta}}{(1+e^{\eta})e^{\beta_0}}\right\}=0. \quad (5)$$

We can show that $U(\eta)$ is an decreasing (increasing) function for any fixed $\beta_0 > 0$ ($\beta_0 < 0$). Now recall that ε_h must satisfy

$$\delta=(1+\varepsilon_h)^{-1/\varepsilon_h}-(1+\varepsilon_h)^{-(1+\varepsilon_h)/\varepsilon_h}, \quad (6)$$

and our true model is POSM of (2.1) with $\beta = \beta_0$ such that $\delta=\frac{|e^{\beta_0/2}-e^{-\beta_0/2}|}{(1+e^{\beta_0/2})(1+e^{-\beta_0/2})}$ (margin for the true maximum of $|S_1(t) - S_0(t)|$). We can numerically show that $U(\eta) < 0$ for $\beta_0 \in (-\infty, 0) \cup (0,+\infty)$. Therefore, we can conclude that when $\beta_0 < 0$, $\eta^* > -\log(1 + \varepsilon_h)$ and when β_0

> 0, $\eta^* < \log(1+\varepsilon_h)$. This is equivalent to $|\eta^*| < \log(1+\varepsilon_h)$ when the true model is POSM with $\beta_0 \neq 0$.

Table 1

For different values of maximum difference in survival curves $m = max|S_1(t) - S_0(t)|$, the Pr(Rejecting H_0) for using POSM based test when the true model is POSM (sample size = $n_1 + n_2$ for $n_1 = n_2$).

Sample Size	*Power*		*Type − I error rate*		
	***m* = 0**	***m* = 0.10**	***m* = 0.15**	***m* = 0.20**	***m* = 0.30**
50	0.114	0.072	0.049	0.030	0.006
100	0.210	0.115	0.050	0.010	0.000
150	0.378	0.154	0.050	0.012	0.000
200	0.598	0.200	0.055	0.007	0.000
400	0.930	0.308	0.044	0.004	0.000

Table 2

For different values of maximum difference in survival curves $m = max|S_1(t) - S_0(t)|$, Pr(Rejecting H_0) for using log-rank based test when the true model is POSM (sample size = $n_1 + n_2$ for $n_1 = n_2$).

Sample Size	*Power*		*Type − I error rate*		
	***m* = 0**	***m* = 0.10**	***m* = 0.15**	***m* = 0.20**	***m* = 0.30**
50	0.120	0.085	0.069	0.032	0.007
100	0.286	0.155	0.085	0.038	0.006
150	0.497	0.235	0.111	0.034	0.004
200	0.685	0.335	0.130	0.033	0.000
300	0.964	0.539	0.180	0.033	0.000

Table 3

For different values of maximum difference in survival curves $m = max|S_1(t) - S_0(t)|$, Pr(Rejecting H_0) for using log-rank based test when the true model is PHM (sample size = $n_1 + n_2$ for $n_1 = n_2$).

Sample Size	*Power*		*Type − I error rate*		
	***m* = 0**	***m* = 0.10**	***m* = 0.15**	***m* = 0.20**	***m* = 0.30**
50	0.127	0.087	0.045	0.012	0.005
100	0.268	0.131	0.052	0.016	0.000
150	0.510	0.156	0.047	0.010	0.000
200	0.676	0.201	0.049	0.009	0.000
400	0.966	0.355	0.052	0.002	0.000

Table 4

For different values of maximum difference in survival curves $m = max|S_1(t) - S_0(t)|$, Pr(Rejecting H_0) for using POSM based test when the true model is PHM (sample size = $n_1 + n_2$ for $n_1 = n_2$).

Sample Size	*Power*		*Type − I error rate*		
	***m* = 0**	***m* = 0.10**	***m* = 0.15**	***m* = 0.20**	***m* = 0.30**
50	0.104	0.076	0.042	0.026	0.008
100	0.211	0.116	0.046	0.015	0.001
150	0.381	0.143	0.044	0.009	0.000
200	0.594	0.207	0.049	0.003	0.000
400	0.920	0.302	0.044	0.003	0.000