Kaplan-Meier based Methods to Address Non-Proportional Hazard Issues

Bo Huang[*]

Abstract

In the clinical development of cancer immunotherapies and targeted therapies, the proportional hazard (PH) assumption used in the power calculation for time-to-event endpoints often does not hold. For example, due to the delayed and (for some patients) durable antitumor effect on cancer cells induced by immunotherapies, the survival curves of a randomized controlled study may take a while to separate and the curve for the immunotherapy agent may have a long and flat tail. Therefore, the log-rank test may lose power, and the interpretation of the hazard ratio (HR) from the standard Cox regression model is not straightforward. Kaplan-meier (KM) based methods such as the restricted mean survival time (RMST) and weighted KM-based tests are interesting alternative methods for statistical inference that do not rely on the PH assumption. The RMST is an appealing statistical measure to quantify treatment benefit in a clinically meaningful and interpretable manner. In this paper, we give an overview of these methods and present a simulation study to compare the performance of the KM-based methods with the HR/log-rank test under various non-PH patterns.

Key Words: Immunotherapy, Efficacy, Non-Proportional Hazard, Delayed Effect, Time-to-Event

1. Introduction

In many disease areas, a time-to-event (TTE) endpoint is used as the primary endpoint in randomized clinical trials. The most common method for the analysis of TTE data is the log-rank test and the Cox proportional-hazard (PH) model under the PH assumption, and the hazard ratio (HR) derived from the Cox-model is used to quantify the relative effect between treatment groups. Under the PH assumption, the log-rank test is the most power test, and the HR can be interpreted approximately as a constant relative measure of risk over time, for instance $S_1(t) = S_0(t)^{HR}$, where $S_1(t)$ and $S_0(t)$ are the the survival function at time t for two randomized treatment groups.

With the emergence of novel therapies such as targeted therapies and immunotherapies in cancer, the PH assumption often does not hold. Depending on the mechanism of the drug, non-PH can demonstrate different patterns. For example, unlike chemotherapies that attach tumor cells directly, immunotherapies target the immune system to elicit effect on cancer cells. The indirect effect could lead to delayed effect as manifested by late separation on the Kaplan-Meier (KM) curves. The delayed treatment effect can also be maintained thanks to its durability on some patients resulting in flat survival tails and long-term benefit. In contrast, a small molecule targeted therapy may manifest fast and dramatic tumor regression early on, however, the effect may not last and could diminish over time, with the survival curves demonstrating a diminishing or "belly-shape" pattern (Figure 1). In both cases, the PH assumption is violated.

The log-rank test is still statistically valid under non-PH, but it may suffer significant power loss owing to its relationship with the score test in the Cox-PH model. The loss in power increases the failure rate of clinical trials for promising novel therapies. Furthermore, the HR is not interpretable under non-PH despite the fact that it can be approximated as the weighted HR over time on the log-scale (Huang and Kuan, 2018).

In the literature, a number of methods have been proposed to tackle the non-PH issue, such as, notably, weighted log-rank test (e.g. Fleming and Harrington, 1981), or a lin-

[*]Pfizer Inc., 445 Eastern Point Rd, Groton, CT 06340

ear combination or the maximum of several test statistics (Tarone, 1981; Gastwirth, 1985; Zucker and Lakatos, 1990; Self, 1991; Lee, 1996). The test proposed by Yang and Prentice (2010) is a weighted logrank test whose weights are obtained by fitting the data to a proposed model.

Since we conventionally present the KM curves to show the temporal profile of survival data, it is natural to perform a test that directly compares two survival function, rather than their hazard functions. Such interesting class of methods is the KM-based method that measures the relative between-group difference by the difference or ratio between the KM curves (Pepe and Fleming, 1989; Royston and Parmar, 2011; Uno et al., 2014), which is different from the rank-based methods (such as the log-rank test or weighted log-rank test) that essentially assess effect by (weighted) difference of hazard functions. Without parametric extrapolation, the KM curves contain all the information of the survival data. Therefore, the KM-based method is an appealing robust approach to summarize between-group difference that does not rely on the PH assumption.

In this paper, we give an overview of three KM methods proposed in the literature and discuss their benefits and limitations. Section 2 describes the concepts of each method. In Section 3, we conduct a simulation study to compare these methods under various non-PH scenarios. Section 4 concludes with a discussion.

2. Kaplan-Meier Based Methods

2.1 Restricted Mean Survival Time (RMST)

The most widely used KM-based method is the restricted mean survival time (RMST), which is a robust and clinically interpretable summary measure of the survival time distribution that does not rely on the PH assumption. Unlike the median survival time, it is estimable even under heavy censoring. There has been considerable methodological research (e.g., Zucker, 1998; Royston and Parmar, 2011; Royston and Parmar, 2013; Uno et al., 2014) on the use of RMST to estimate treatment effects as an alternative to the HR approach. The RMST methodology is applicable independent of the PH assumption, and a test of the difference or ratio between the RMST for the experimental arm and the control arm may be more appropriate to determine superiority with respect to the time-to-event endpoint. The RMST depends on the selection of cutoff (truncation) time τ, which needs to be pre-specified to avoid selection bias (after seeing the data). Common selections include fixed landmark times of clinical relevance (e.g. x-year), minimum of the largest observed event time in each of the two groups, or minimum of the largest observed time (event or censoring) in each of the two groups.

The RMST μ of a random time-to-event variable T is the mean of the survival time $X = \min(T, \tau)$ truncated at a cutoff time $\tau > 0$. It can be derived as the area under the survival curve $S(t) = P(T > t)$ from $t = 0$ to $t = \tau$:

$$\mu(\tau) = E(X) = \int_0^\tau S(t)dt \tag{1}$$

The variance term $\sigma^2(\tau)$ of X can also be derived accordingly using integration by part:

$$\sigma^2(\tau) = \text{Var}(X) = 2\int_0^\tau tS(t)dt - \left[\int_0^\tau S(t)dt\right]^2 \tag{2}$$

A natural estimator for μ is

$$\hat{\mu}(\tau) = \int_0^\tau \hat{S}(t)dt \tag{3}$$

where $\hat{S}(t)$ is the KM estimator for the survival function of T, a step function with mass at the time points $t_1, t_2, \ldots, t_D$. $\hat{\mu}(\tau)$ approximately follows a normal distribution with its variance term estimated below:

$$V[\hat{\mu}(\tau)] = \sum_{i=1}^{D} \left[\int_{t_i}^{\tau} \hat{S}(t)dt \right]^2 \frac{d_i}{Y_i(Y_i - d_i)} \tag{4}$$

where d_i and Y_i are the number of events and number of patients at risk at t_i, respectively.

In a randomized two-arm trial with survival function $S_T(t)$ and $S_C(t)$ for the treatment arm and control arm, respectively, the difference in RMST between arms can be estimated as

$$\int_0^{\tau} [\hat{S}_T(t) - \hat{S}_C(t)]dt \tag{5}$$

with estimated variance term $V[\hat{\mu}_T(\tau)] + V[\hat{\mu}_C(\tau)]$.

Alternatively, analogous to the hazard ratio as a measurement of the relative risk of event hazard, a similar measurement for RMST is the ratio of RMST between arms (control versus treatment), with ratio < 1 indicating survival improvement in the treatment arm. Unlike the HR, the RMST ratio does not rely on any model assumption, which can be estimated as

$$\frac{\int_0^{\tau} \hat{S}_C(t)dt}{\int_0^{\tau} \hat{S}_T(t)dt} \tag{6}$$

with variance term estimated using the delta method.

2.2 Weighted Kaplan-Meier test (Pepe and Fleming, 1989)

Pepe and Fleming (1989) proposed a class of tests based on the weighted KM (WKM) statistic. The test is based on a linear combination of weighted differences of 2 KM curves over time and when the weight function is a constant 1, the WKM statistic is equivalent to the RMST difference. However, this method did not get much attention in practice.

Specifically, let $\hat{S}_T(\cdot)$ and $\hat{S}_C(\cdot)$ be the KM estimators for the treatment and control groups to be compared. A WKM test statistic is

$$\left(\frac{n_T n_C}{n_T + n_C}\right)^{1/2} \int_0^{\tau} \hat{W}(t)\hat{D}(t)dt \tag{7}$$

where $\hat{D}(t) = \hat{S}_T(t) - \hat{S}_C(t)$, $\tau = \sup\left[t : \min\left\{\hat{K}_T(t), \hat{K}_C(t)\right\} > 0\right]$, $\hat{K}_i(\cdot)$ denotes the left-continuous version of the KM estimator for the censoring survival function for Group i, n_i is the sample size in Group i ($i = T, C$), and $\hat{W}(\cdot)$ is the weight function.

For the test statistics in (7), Pepe and Fleming (1989) proposed two weighting schemes:

$$\frac{\hat{K}_T(t)\hat{K}_C(t)}{\hat{q}_T\hat{K}_T(t) + \hat{q}_C\hat{K}_C(t)} \tag{8}$$

and

$$\left\{\frac{\hat{K}_T(t)\hat{K}_C(t)}{\hat{q}_T\hat{K}_T(t) + \hat{q}_C\hat{K}_C(t)}\right\}^{1/2} \tag{9}$$

where $\hat{q}_i$ is the percentage of patients assigned to Group i. The weighting schemes are essentially functions of inverse probability of censoring and put more weight on early time points and less weight on late time points. The objective of such weighting schemes is to increase the stability of the test statistics by assigning less weight to the tails of the KM curves, analogous to the Wilconxon test or Fleming-Harrington test $FH(p, 0)$ where $p > 0$.

2.3 Weighted Kaplan-Meier test (Uno et al., 2014)

Uno et al. (2014) proposed a data-dependent weight function that automatically makes weighting adjustment, with the weight at study time t proportional to $\hat{D}(t) = \hat{S}_T(t) - \hat{S}_C(t)$. Let $\hat{\sigma}(\cdot)$ be the standard error estimate of $\hat{D}(t)$, and $Z(\cdot) = \hat{D}(\cdot)/\hat{\sigma}(\cdot)$, which is distributed approximately $N(0, 1)$ under the null hypothesis.

Instead of utilizing $Z(t)$ as a test statistic, Uno et al. (2014) consider a test statistic which is a weighted integration of standardized differences between two survival curves over $[0, \tau]$

$$V = \int_0^\tau \hat{W}(t)Z(t)dt \tag{10}$$

where $\hat{W}(\cdot)$ is a data-dependent weight function. One proposed weighting scheme is $\hat{W}_c(t) = \max\{Z(t), c\}$ and

$$V_1(c) = \int_0^\tau \hat{W}_c(t)Z(t)dt \tag{11}$$

and c is selected adaptively to construct a test statistic based on $\{V_1(c), 0 \leq c \leq \eta\}$, where η is a constant.

The test automatically makes weighting adjustments empirically by putting more weight when "difference" is large. The proposed test is data-driven and agnostic to various non-PH patterns. Similar to the method by Yang and Prentice (2010), Uno and colleagues warned that the Type I error rate might be slightly inflated when the sample size or the number of observed events is small.

One disadvantage of this method is that it is computationally very intensive because it employs the perturbation resampling approach to approximate the distribution under the null hypothesis.

3. Simulation Study

In this simulation study, survival time is assumed to follow a piecewise exponential distribution with piecewise constant hazard in each time interval defined by a sequence of change-points. Patients are $1 : 1$ randomized to the treatment arm and the control arm. We further assume that the drop-out censoring variable follows an exponential distribution with hazard rate of 0.004 in both arms.

Scenario 1 assigns hazard rates of 0.104 and 0.103 to the treatment and control arms respectively for the first 3 months, and hazard rates of 0.161 and 0.077 afterwards. In other words, a HR of approximately 1 in the first 3 months and HR of 0.48 afterwards (3-month delayed effect with PH afterwards). The analysis time for all methods to be evaluates is when 70% of a total of 300 patients have events. Scenario 2 assumes the control arm has a constant hazard rate of 0.069 (median survival time of 10 months), while the treatment arm hazard function is piecewise constant with values of 0.069, 0.052 and 0.001 in the first 5 months, 5 to 15 months and after 15 months respectively (corresponding to a HR of $1, 0.75$ and 0.02 in each interval), indicating a 5-month delayed effect with long-term survival pattern. The analysis time for all methods to be evaluates is when 70% of a total of 300 patients have events. Under Scenario 3, treatment effect is diminishing over time, with a constant hazard rate of 0.069 in the control arm, and piecewise constant hazards of 0.045 and 0.083 for the treatment arm in the first 15 months and after 15 months. The analysis time for all methods to be evaluates is when 70% of a total of 300 patients have events. Scenario 4 assumes a 3-month delay with long-term survival and crossing hazard pattern, with control arm hazard rates equal to 0.347, 0.116, 0.046, and treatment arm hazard rates equal to 0.520, 0.116, 0.007 in the first month, 1 month to 3 months and after 3 months respectively. The analysis time for all methods to be evaluates is when 75% of a total of 300 patients have events.

Accrual is assumed to follow a ramp-up pattern with slow enrollment in the first 6 months and faster and constant accrual rate afterwards.

For each scenario, 5000 simulations are performed to evaluate the performance of each method. The results are summarized in Table 1).

Table 1: Simulation results under Scenarios 1-4 for the comparison of the log-rank test, RMST test, WKM test (Pepe and Fleming, 1989) and WKM test (Uno et al., 2014). HR and RMST difference are summarized under each scenario. The log-rank test and the HR use data up to the minimum of (maximum of the largest observed event time in either arm, minimum of the largest event/censoring time in either arm). The KM-based methods use data up to the minimum of the largest event/censoring time in either arm.

	Log-rank		RMST		WKM (PF)	WKM (Uno)
	HR	Power	Diff.(m)	Power	Power	Power
Sc1 (3m delay then PH)	0,64	89.7%	2.52	90.3%	78.3%	90.9%
Sc2 (5m delay with long-term survival)	0.78	46.9%	3.01	53.3%	28.8%	68.4%
Sc3 (diminishing, crossing hazards)	0.75	57.1%	2.65	56.0%	68.5%	60.7%
Sc4 (3m delay with long-term survival, crossing hazards)	0.78	48.9%	5.36	68.6%	44.1%	92.1%

The log-rank test performs well in Scenario 1 when there is a 3-month delayed effect and PH afterwards. Since the majority of events occur after 3 months, the loss in power due to non-PH is controlled to a less extent, with an overall estimated HR of 0.64 from the Cox-PH model. RMST test and WKM test (Uno) perform similarly, with the WKM test (Uno) having the highest power. The WKM test (PF) does not perform well due to assigning less weight to longer follow-up. In Scenario2 and Scenarios 3 when there is a delayed effect (3-5 months) and the delayed effect is maintained with the treatment arm KM curve flat in the tails, indicating a long-term survival or remission pattern (sometime referred to as mixture cure-rate survival model), the log-rank test has substantial power loss, while the RMST test and the WKM (Uno) test have much higher power. The WKM (PF) test does not perform well. In Scenario 3 with a diminishing and crossing-hazard pattern, the WKM (PF) test has the best performance, followed by the WKM (Uno) test. The log-rank test and the RMST test have similar power. Overall, the WKM (Uno) test has the highest power across all 4 different scenarios, followed by the RMST test.

4. Discussion

The KM-based method is an appealing approach to address the issue of non-PH for the analysis of survival data, in particular for a drug that may demonstrate long-term benefit.

Table 2 summarizes the pros and cons of the KM-based method in comparison with the HR and log-rank test. The WKM test (Uno) utilizes a data-driven weighting scheme and can handle all types of non-PH scenarios. Simulations show that it has the highest power of compared to the log-rank test and other KM-based tests. However, it is computationally intensive and relies on a resampling approach to control the Type I error rate. The WLM test (PF) has poor performance except for the diminishing effect scenario. The test based

Table 2: Pros and Cons of KM-based methods versus the HR/log-rank test in the presence of non-PH.

	HR and log-rank test	RMST	WKM test (PF)	WKM test (Uno)
Power	may have substantial power loss	higher power than the log-rank in some scenarios	poor performance except for diminishing effect	robust and high performance in all scenarios
Sensitive to long term survival or remission	no	yes	no	yes
Robust to all non-PH types	no	no	no	yes
Estimation	HR	RMST difference RMST ratio	no estimation measure	no estimation measure
Clinical interpretation under non-PH	clinically not interpretable	simple and meaningful interpretation	weighted test difficult to interpret	weighted test difficult to interpret
Asymptotics	yes	yes	yes	no (resampling)
Computing speed	fast	fast	fast	slow

on RMST can have higher power than the log-rank test in some non-PH scenarios. Both the RMST test and the WKM test (Uno) are sensitive to long-term survival or long-term remission. The log-rank test can have substantial power loss in the presence of non-PH.

Other than hypothesis testing, another important aspect of statistical inference is the estimation. When the PH assumption is violated, the HR derived from the Cox-PH model is not clinically interpretable and the difference in median may under-estimate or over-estimate the treatment effect. The WKM tests are testing approaches without associated estimation measures. The RMST is a clinically and statistically meaningful global summary measure no matter if the PH assumption holds. Figure 2 illustrates the advantage of using RMST difference to measure treatment benefit in comparing inotuzumab with standard chemotherapy in a Phase 3 randomized study for patients with relapsed or refractory acute lymphoblastic leukemia (Kantarjian et al., 2016). We recommend that the RMST estimation and testing should be included as a regular analytic procedure in the toolkit of survival analysis.

REFERENCES

Fleming TR, Harrington DP. Counting processes and survival analysis. Wiley, 1991.

Tarone RE. On the distribution of the maximum of the logrank statistic and the modified Wilcoxon statistic. *Biometrics* 1981; **37**:79-85.

Gastwirth JL. The Use of Maximin Efficiency Robust Tests in Combining Contingency Tables and Survival Analysis. *Journal of the American Statistical Association* 1985; **80(390)**:380-384.

Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* 1990; **77(4)**:853-864.

Self SG. An adaptive weighted log-rank test with application to cancer prevention and screening trials. *Biometrics* 1991; **47(3)**:975-986.

Lee JW. Some versatile tests based on the simultaneous use of weighted log-rank statistics. *Biometrics* 1996; **52**:721-725.

Yang S, Prentice R. Improved logrank-type tests for survival data using adaptive weights. *Biometrics* 2010; **66(1)**:30-38.

Zucker D. Restricted mean life with covariates: modification and extension of a useful survival analysis method. *Journal of the American Statistical Association* 1998; **93**:702-709.

Royston P, Parmar M. The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine* 2011;

30(19):2409-2421.

Royston P, Parmar M. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC medical research methodology* 2013; **13(1)**:152-166.

Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics* 1989; **45**:497-507.

Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, ... Wei LJ. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. *Journal of Clinical Oncology* 2014; **32(22)**:2380-2385.

Kantarjian H, DeAngelo D, Stelljes M, Martinelli G, Liedtke M, Stock W, Gokbuget N, OBrien S, Wang K, Wang T, Paccagnella L. Inotuzumab ozogamicin versus standard therapy for acute lymphoblastic leukemia. *New England Journal of Medicine* 2016; **375(8)**:740-753.

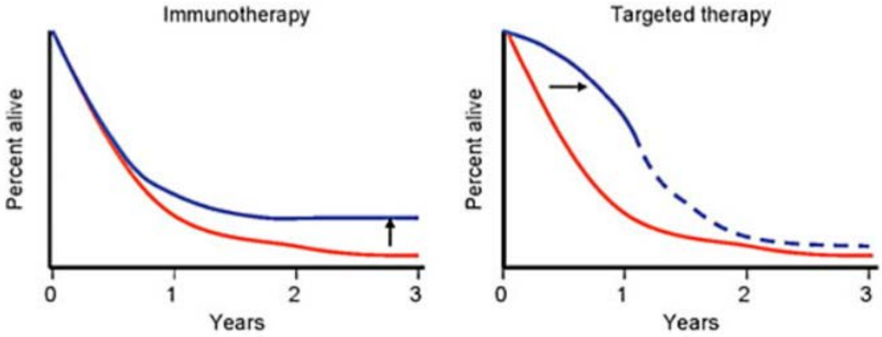

Figure 1: Examples of patterns of non-proportional hazards based on mechanisms of actions.

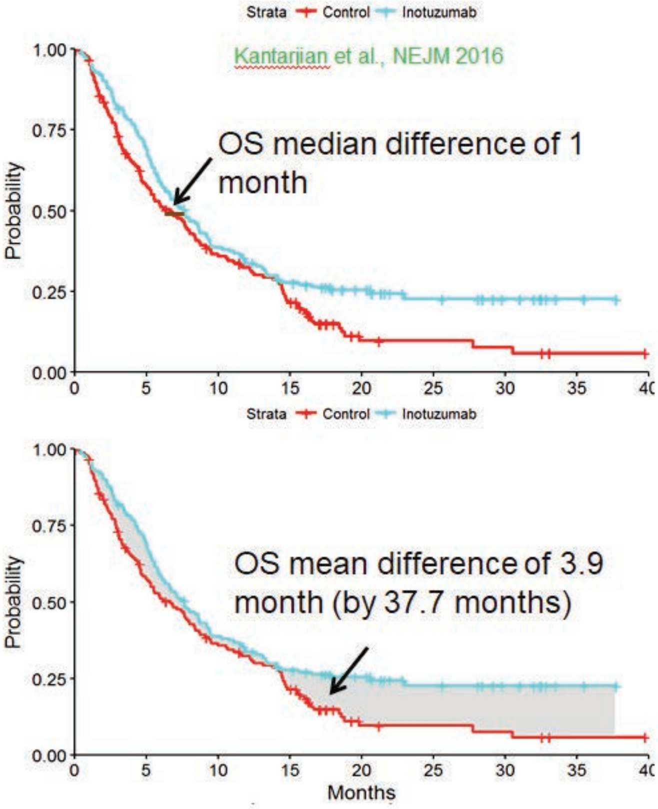

Figure 2: Overall survival Kaplan-Meier curves of the phase 3 randomized study in patients with relapsed or refractory, CD22-positive, Philadelphia chromosome (Ph)-positive or Ph-negative acute lymphoblastic leukemia. A total of 326 patients were 1:1 randomized to receive either inotuzumab ozogamicin (inotuzumab ozogamicin group) or standard intensive chemotherapy (standard-therapy group) (Source: Kantarjian et al., 2016).

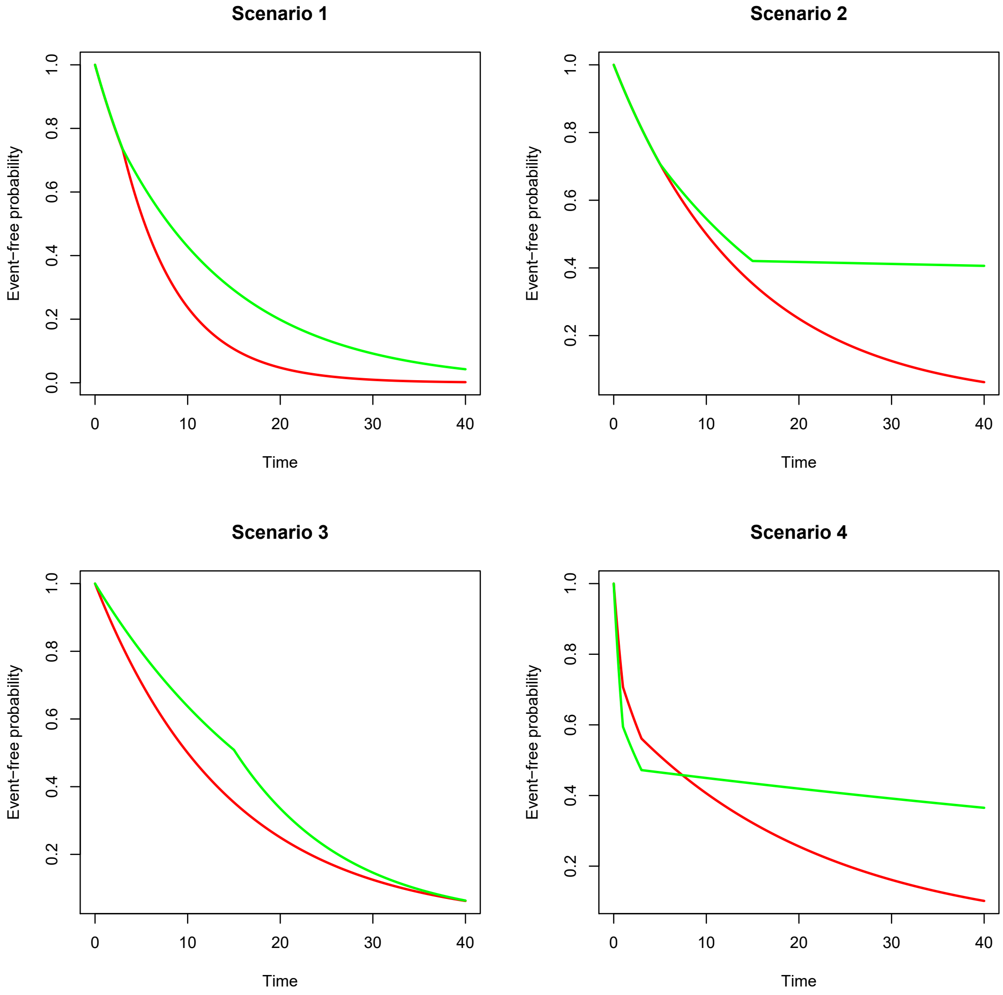

Figure 3: Non-PH simulation scenarios of event-free probability by time for a randomized $(1 : 1)$ clinical trial comparing the treatment arm (green) with the control arm (red).