**Paper PP16**

# Comparison of Standard Log-Rank Test with Selected Alternatives

Elena Valkanova, ICON PLC
Bulat Fatykhov, ICON PLC

## ABSTRACT

In oncology studies with time-to-event outcomes, the hazard and survival functions are two ways of determining survival distribution; comparing the survival distributions using the constant hazards assumption could provide misleading conclusions. The clinical trial designs with two groups under the proportional hazards (PH) assumption is likely to be underpowered to detect treatment effects. When the PH assumption is violated, the likelihood estimator depends on the underlying censoring time distribution.
The objective of this paper is comparing selected alternative methods for survival analysis when the assumption for the proportional hazards is questionable. We discussed three selected methods: log-rank (LR), Max-Combo (MC) and restricted mean survival time (RMST) applied to simulation data and real clinical trial data supported with graphical presentations using R. In the final section we provide guidance in choosing an alternative method to the standard log-rank test.

## 1. INTRODUCTION

In oncology randomized controlled trials (RCTs) with two treatment groups, it is of fundamental importance choosing statistical methods that infer a difference in the survival distributions with censored time-to-event data. The standard log-rank (LR) test is commonly used to compare the overall survival of two groups in balanced superiority trials, and it is the most powerful and fully efficient test under the assumption of proportional hazards (PH).

Testing the PH assumption is critical, and an extended study on pooled 175 oncology trials indicated in only 31% of these studies the assumption was tested when 47% of all studies had crossing curves (Kristiansen, 2012).

In trials when PH assumption is violated, the power of a log-rank test tends to decrease, and might generate incorrect results in small or medium sample sizes (Qiu, 2008; Li, 2015). In many immunotherapy oncology and transplant clinical trials, the PH assumption does not hold due to delayed treatment effect. In the case of anticipating an early treatment effect, the RMST test is preferred since it has more power than the log-rank test (Tian, 2018).

The discussion of alternative methods was initiated in a public workshop by the FDA in 2018, which proposed a statistical test sometimes referred to as the "Max-Combo" (MC) (Duke, 2018). The MC test is one of the robust methods if no preliminary information about the trial is available, and it handles different types of treatment effects, such as delayed and early effects (Lin, 2020; Roychoudhury, 2021).

It has been reported that crossing survival curves are expected when one of the treatments offers a short-term benefit but no long-term advantages, or when the treatments being compared have different biological mechanisms of action (Royston, 2013). Furthermore, the combined methods performed better in some cases of uncertainty about the survival time

distributions or PH violations (Dormuth, 2023). Different tests have been conducted on the reconstructed individual patient data, and Fleming-Harrington (F-H) log-rank tests with weight functions have been recommended to increase sensitivity for detecting drug effects on survival in cancer vaccine or immunotherapy oncology trials (Castañon, 2020).

In a recent systematic review, which evaluated the performance of the log-rank with the Max-Combo and RMST applied on 63 immuno-oncology completed trials, including 35,902 patients, it was shown that in 10% of the selected trials, the Max-Combo test achieved higher significance vs. the log-rank test. Hence, several studies showed that treatment had clinically meaningful benefits, but it would not have been detected using a log-rank test (Mukhopadhyay, 2022).

In our paper, we simulated data for 1,000 studies covering the three scenarios: crossing effect, late effect, and early effect, by controlling for different parameters using the Weibull and Exponential distributions. The main goal of this paper is to compare the power of the three tests, log-rank vs. MC and RMST, for small and medium sample sizes since the tests produce similar power for large sample sizes. There are more alternative tests to LR, but we are analyzing the two alternative tests, MC and RMST, as the most preferred methods due to their robust performance and easy interpretation.

Our study has used simulated data along with the reconstructed data to compare the power performance of the tests by controlling for Type I errors. The conclusion from the results aligned with the latest study by Dormuth et. al. (2023).

The article's structure: Section 2 covers necessary notations and definitions used in the methods. In Section 3, we describe the simulation methods used to evaluate the performance of the MC and RMST and compare all three methods by comparing power and Type I errors. In Section 4, we provide case study results from "Fulvestrant plus capivasertib versus placebo after relapse or progression on an aromatase inhibitor in metastatic, oestrogen receptor-positive breast cancer (FAKTION): a multicentre, randomized, controlled, phase 2 trial" (ClinicalTrials.gov, NCT01992952) (Jones, 2020). The FAKTION trial investigated whether the addition of capivasertib to fulvestrant improved progression-free survival and overall survival in patients with advanced breast cancer.

Section 5 discusses the limitations and guidelines for choosing an alternative method to the standard log-rank test.

## 2. METHODS

### Assumptions and Notations

The parameters below were controlled during the simulation study, and different settings and results are included in Table 1, Table 2, and Table 3 (Appendix).

- $\mathbf{P_{censoring}}$ - probability of overall survival (OS).
- **N** - the number of subjects in each group or total population is 2N.
- $\mathbf{t_{effect}}$ - effect time indicates the time at which treatment is kicking off. The hazards in the treatment and control groups are the same as before $t_{effect}$.

- **$dur_{effect}$** - effect duration indicates the duration of the treatment effect. The hazards in the treatment and control groups remain the same after $dur_{effect}$.

The assumptions used in the simulation study are:

1) Survival times for an event of interest and censoring times are independently and identically distributed (iid) according to the cumulative distribution function and survival function, respectively.

2) When the generated survival time is greater than the trial duration, it is set equal to the trial duration and censored at the end of the trial.

The data sets generated covered the following three scenarios: crossing curves (CC) and non-crossing curves (late effect and early effect). In the CC scenario, we generated the survival time of the treatment groups based on the Weibull distribution with a fixed scale parameter of 1.4 and a varying shape parameter that ranges from 1 (the PH assumption is met) to 4 (the PH assumption is violated). In the late effect and early effect scenarios, the survival time of the treatment group was generated from the combination of exponential distributions to ensure a lower hazard rate in the treatment group from the beginning of the study (early effect) or from a particular point of the study (late effect). We generated data sets for each scenario with varying sample sizes (50, 100, or 200 subjects in each treatment group), treatment durations (1, 2, or 3), and different censoring probabilities ranging from 0.2 to 0.4. Considering every scenario and combination of four parameters, we simulated data for 1,000 studies and estimated the power of each of the three tests. In addition, the utility of the three tests is further evaluated on the individual patient data reconstructed from a FAKTION study (2020): a multicenter, randomized, controlled, phase 2 trial using the algorithm introduced by Guyot et al. (Guyot, 2012).

## 2.1 Maximum Weighted Log-rank Tests (Max-Combo)

The maximum weighted log-rank tests (Max-Combo) are based on combining multiple weighted log-rank tests and multiple test statistics with different weights, where the final test statistics are defined on the maximum of all. The MC test could be applied to the three scenarios of early effect, late effect, and similar effect. However, it requires preliminary knowledge of specific therapeutic areas to properly allocate the weights to the treatment groups. The main advantage of MC is that it improves power under a wide range of non-PH patterns and outperforms the log-rank test.

## 2.2 Restricted Mean Survival Times (RMST)

The RMST method is classified under the methods for areas under the survival curve, and it can be used for either superiority or non-inferiority studies of balanced group sizes with crossing hazard patterns. The RMST has been proposed as an alternative measure to overcome some of the limitations of proportional hazard modeling (Royston & Parmar, 2011, 2013) in medical fields such as oncology (Uno, 2014; Trinquart, 2016; Pak, 2017), pulmonary medicine (Harhay, 2018), and cardiology (Kim, 2017; Perego, 2020). A comparison study of hazard ratios vs RMST ratios by Trinquart et al. indicated that treatment-effect measures based on

RMST yield more conservative estimates than HRs in oncology randomized controlled trials (RCTs) (Trinquart, 2016). The advantages of the RMST method are that it is independent of distribution and makes it easy to clinically interpret the primary outcome results. The challenge of the method is that an explicitly chosen time point is required to obtain an RMST associated with the clinically relevant time horizon (Han & Jung, 2022).

The restricted mean survival time of a random variable T, is the mean of the survival time defined as: $X = \min(T, t^*)$, where $t^*$ is a pre-specified time horizon.

$$\mu = E(X) = E[\min(T, t^*)] = \int_0^{t^*} S(t)dt.$$

It is the area under the survival curve $S(t) = P(T > t)$ from $t = 0$ *to* $t = t^*$. In the two-group clinical trial with survival functions: $S_0(t)$ and $S_1(t)$, the area between the survival curves calculated by:

$$\Delta = \int_0^{t^*} S_1(t) - \int_0^{t^*} S_0(t) = \int_0^{t^*} [S_1(t) - S_0(t)]dt$$

## 2.3 Example

Figure 1 displays an example of different scenarios of survival curves for two treatment groups:

a) PH effect, b) Non-PH/Early effect, c) Non-PH/Late effect, and d) Non-PH/Crossing effect.

Figure 1. KM Plots for different treatment effects [Royston and Parmar (2020)].

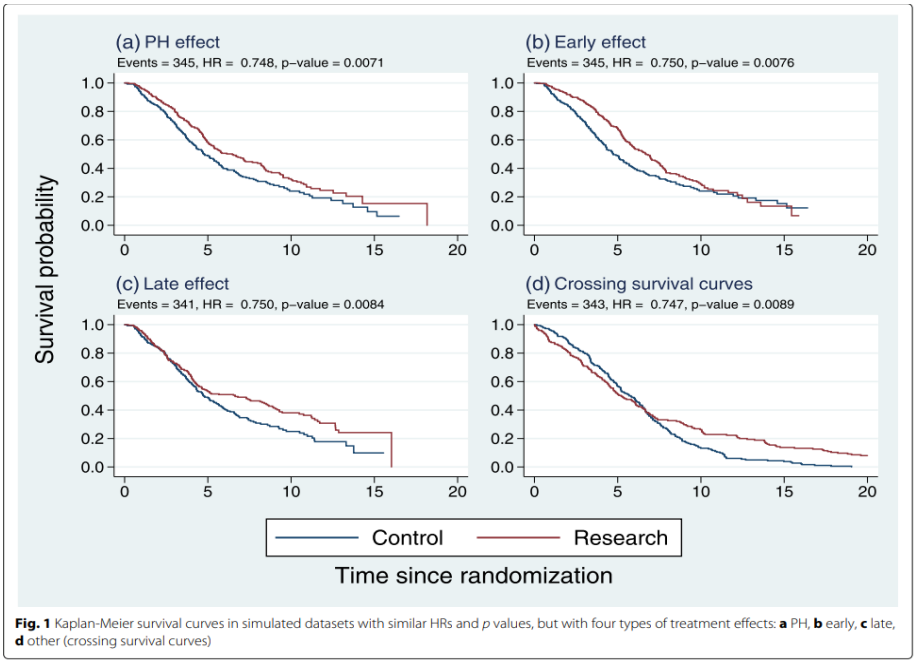



**Fig. 1** Kaplan-Meier survival curves in simulated datasets with similar HRs and *p* values, but with four types of treatment effects: **a** PH, **b** early, **c** late, **d** other (crossing survival curves)

# 3. SIMULATION

To evaluate the performance of the presented methods, we employed simulations for different scenarios and settings using the open-source software R. We simulated data for two groups under exponential and Weibull distributions. Hence, we followed well-established recommendations on the choice of survival distributions.

## 3.1 Data generation process

We simulated data for an oncology balanced study with two treatment groups: a control group and a treatment group, each with N subjects. In the control group, the survival time for each subject is generated from an exponential distribution with a rate of 1. In the treatment group, the survival time for each subject is generated from either a Weibull distribution or a combination of exponential distributions, depending on the specific scenario (see additional details in the next section). The goal is to use simulated data from 1,000 studies to compare the power of each of the three tests: log-rank, Max-Combo, and RMST.
The duration of the study is limited, with three possible values of 1, 2, and 3. For subjects whose survival time is less than the study duration, their overall survival time is censored with a probability of $P_{censoring}$. The subjects who survived after the study duration, their overall survival time is censored at that point.

## 3.2 Scenarios

Data sets were generated for the three scenarios: crossing curves, late and early effects. For each given scenario, we generated data sets with varying sample sizes (50, 100, or 200 subjects in each treatment group), treatment durations (1, 2, or 3), different censoring probabilities ($P_{censoring}$) ranging from 0.2 to 0.4, and parameter that controls the extent to which the PH assumption is violated.

In the scenario of crossing curves, the degree of violation is determined by the shape parameter of the Weibull distribution, as shown in Table 1 (Appendix). The delayed effect scenario is determined by the effect time ($t_{effect}$) parameter, and the early effect scenario is determined by the effect duration ($dur_{effect}$) parameter as shown in Table 2 and Table 3 (Appendix).

- In the Crossing Curves (CC) scenario, the survival time of the treatment group was generated from a Weibull distribution with a scale parameter of 1.4. The shape parameter for the distribution ranges from 1 (Weibull distribution is the same as exponential distribution and the PH assumption is met) to 4 (PH assumption is violated).

- In the late effect (LE) scenario, data generation was controlled by a parameter called $t_{effect}$, which represents the time a hypothetical drug begins to exert its effect, thereby reducing the hazard. The parameter $t_{effect}$ ranges from 0 (indicating that the drug effect starts immediately and there are no late effects) to 0.6 (indicating that the treatment effect starts relatively far from the treatment start date).

For each subject in the treatment group, a value, denoted as t1 is generated from an exponential distribution with a rate of 1:

1. If t1 is less than $t_{effect}$ (indicating that the hypothetical drug had not yet started to work when the subject died), t1 was considered the survival time for that subject.
2. If t1 is greater than or equal to $t_{effect}$ (indicating that the subject survived until the hypothetical drug started to work), a second value, denoted as t2, was generated from an exponential distribution with a rate of 2. The sum of $t_{effect}$ and t2 was used as the survival time for that subject.

- In the early effect (EE) scenario, the data generation process was controlled by a parameter called $dur_{effect}$. This parameter represents the duration, starting from the initiation of treatment, in which a hypothetical drug exerts its effect, leading to a reduction in hazards. The $dur_{effect}$ parameter ranges from 0.3 (representing a short duration of the drug's effect) to 1 (indicating a long-lasting effect of the drug).

For each subject in the treatment group, a value, denoted as t1 is generated from an exponential distribution with a rate of 2.5:

1. If the value t1 is smaller than the $dur_{effect}$ parameter, it means that the hypothetical drug still exerts its effect on the subject at the time of its death or censorship. In such cases, t1 was considered the survival time for that subject.
2. On the other hand, if t1 is greater than or equal to the $dur_{effect}$ parameter, it indicates that the subject survived until the point when the hypothetical drug's effect ceased. In these instances, a second value denoted as t2 was generated from an exponential distribution with a rate of 1. The sum of $dur_{effect}$ and t2 was used as the survival time for that subject.

## 3.3 Power estimation

We considered three main scenarios: crossing curves (CC), late effects (LE), and early effects (EE). For each scenario, we vary the group sizes (50, 100, and 200), trial durations (1, 2, and 3), and censoring rate (from 0.2 to 0.4 by 0.05). In addition, we vary the following:

1. For CC scenario: shape parameter of Weibull distribution (from 1 to 4 by 0.1)
2. For LE scenario: effect time parameter (from 0 to 0.6 by 0.05)
3. For EE scenario: effect duration parameter (from 0.3 to 1 by 0.05)

Hence, we generated 2,655 different settings based on 3 (sample sizes) × 5 (censoring rates) × 3 (trial durations) × [31 (shapes for crossing curves) + 13 (effect times for late effect) + 15 (effect durations for early effect). For each setting, 1,000 replications were performed. The power of each test was subsequently determined by the proportion of studies in which a significant difference in survival functions was detected.

## 3.4 Results

**Type 1 error**

By simulation, we estimated the empirical type I error of the three tests at the $\alpha = 0.05$ level. For each censoring rate, sample size, and trial duration, we simulated 1,000 replicates. Empirical power was measured as the proportion of trials for which a test would have identified a significant result (p-value < 0.05). The results are included in Table 4.

**Non-PH and non-crossing cases:**

Depending on the setting and controlling for different parameters for each scenario, the power of the tests varies significantly. From our simulation results in Table 2 and Figure 2, the log-rank test still has high power for multiple parameter settings and high censoring. At the same time, it no longer dominates all parameter combinations. Hence, the LR test can still be used under moderate violation of the PH assumption, while the MC is more robust against power loss.

**Non-PH and crossing cases:**

In the context of the crossing curves scenario (Table 1), it is apparent that the Max-Combo and RMST tests have more power than compared to the standard log-rank test when the shape parameter has larger values (i.e., when hazards are far from being proportional). As the shape parameter increases, the level of superiority becomes more pronounced. Moreover, the Max-Combo test has more statistical power than the RMST test, especially for larger values of the parameter trial duration with low censoring rate when the survival curves intersect.

# 4. CASE STUDY

We have reconstructed individual patient data from the FAKTION trial (ClinicalTrials.gov, NCT01992952) (Jones, 2020), which investigated whether the addition of capivasertib to fulvestrant improved progression-free survival and overall survival in patients with advanced breast cancer. The quality of the reconstructed data (Figure 6) is sufficiently satisfactory vs. the original KM plot (Figure 5), as HR = 0.61 (95% CI 0.34-1.06), p = 0.082, and HR = 0.59 (95% CI 0.34-1.05), p = 0.071, respectively. Two survival curves cross early time with a clear difference at later times. The two alternative tests were applied to the reconstructed data and compared with the log-rank test.

# 5. DISCUSSION

Our simulation results confirmed the conclusion by Li (2015) and Dormuth (2023) that the Max-Combo test performs much better than the standard log-rank test in the case of the late crossing of the survival curves, and the RMST test performs better in the case of the early crossing of the survival curves.

The empirical power assessment with the simulation study and breast cancer randomized controlled trial provided new insights into the performance of alternative statistical tests beyond the standard log-rank test.

The main advantage of this study is that the analyses are easy to replicate with manageable computing resources, and the R code is available on GitHub.
However, it is important to acknowledge certain limitations of our study. Firstly, our analyses of simulated data are controlled for specific parameters; further evaluations of the characteristics of the proposed methods could facilitate making the correct decision when we anticipate non-PH. Secondly, the trial duration can be extended for a long follow-up period, which could require a sequential testing approach. For example, conduct a standard log-rank test in the first stage, and if the test does not reject the null hypothesis, then asymptotic independent tests for crossing curves are suggested by Qiu (2008). Furthermore, this approach can be extended to multi-stage testing of additional asymptotically independent tests.

By acknowledging these limitations, we ensure transparency in our study and provide a future direction for investigating the characteristics of the alternative methods under different non-PH patterns. Moreover, factors other than power should be taken under consideration when selecting statistical tests for oncology randomized controlled trials.

## DATA SHARING:

Data is available upon request; R the code is posted under:
https://github.com/Fa-bula/RMST_CWLG


## ACKNOWLEDGMENTS

We thank Aisha Vasquez Jackson for reviewing the paper.


## CONTACT INFORMATION:


Elena.Valkanova@iconplc.com
Bulat.Fatykhov@iconplc.com
ICON PLC.

# APPENDIX

## FIGURES

Figure 2. Power of Log-Rank vs MC vs RMST for Crossing Curves ( N=50)

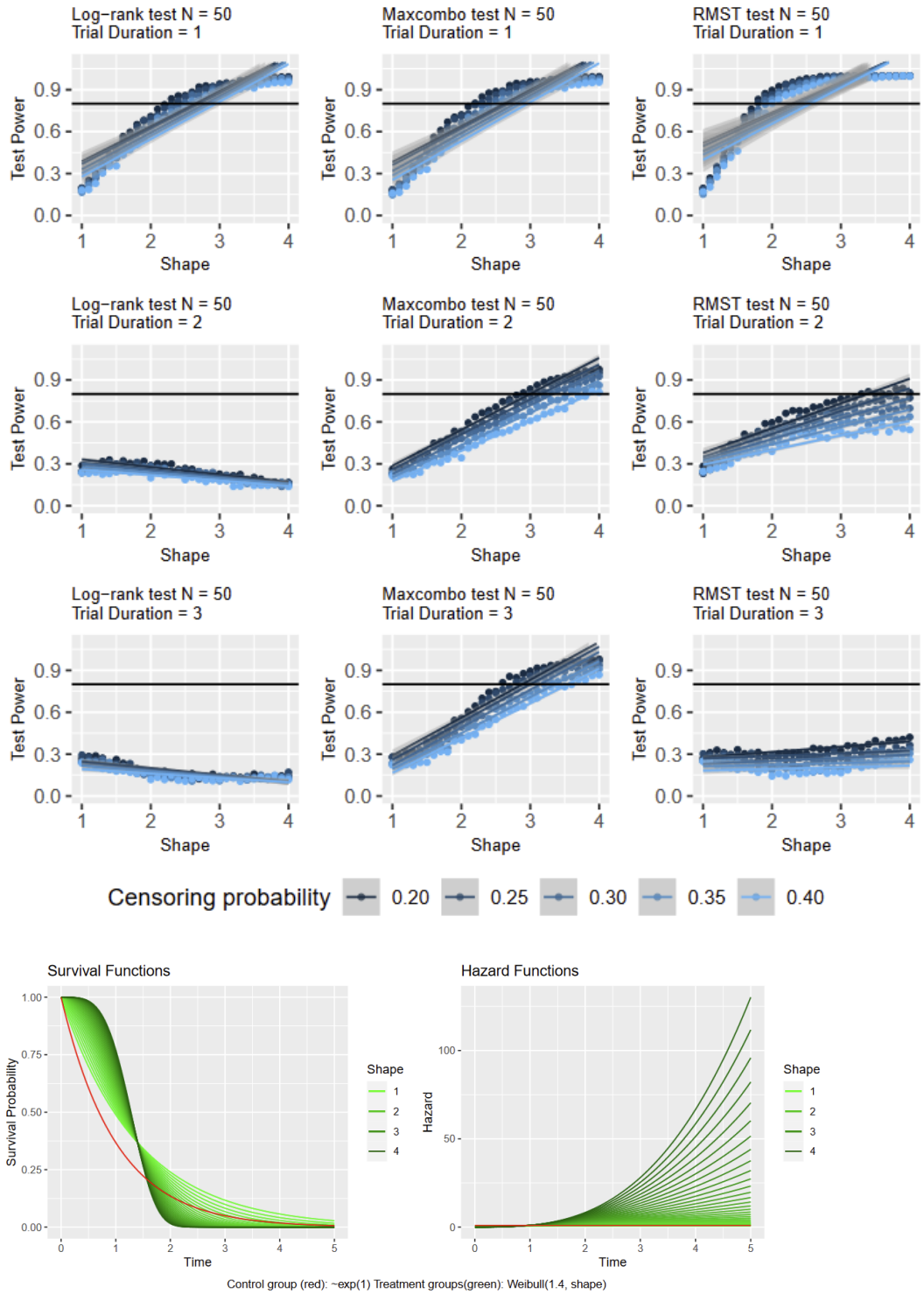


Figure 3. Power of Log-Rank vs MC vs RMST for Early Effect (N=50)

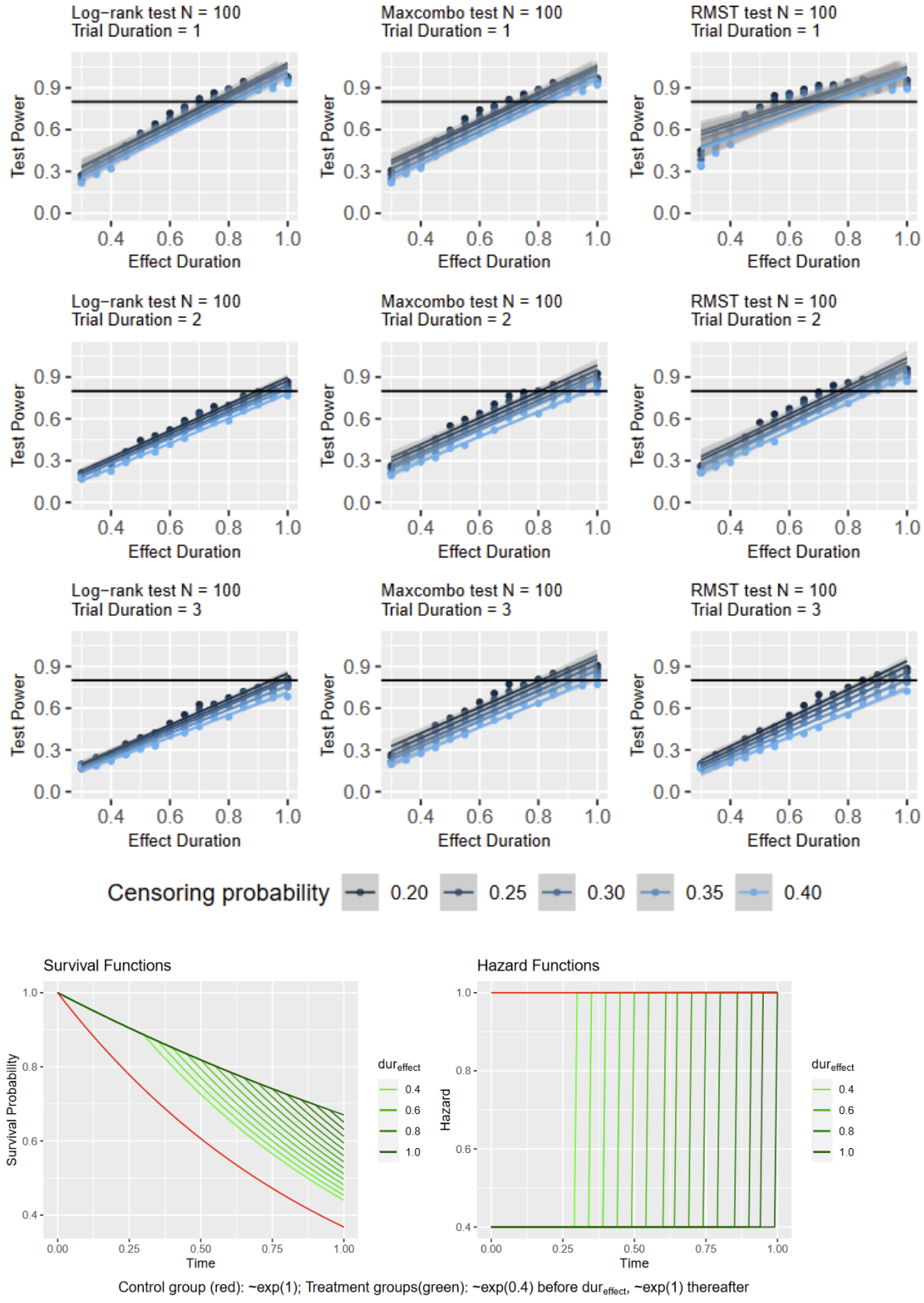


Figure 4. Power of Log-Rank vs MC vs RMST for Late Effect (N=50)

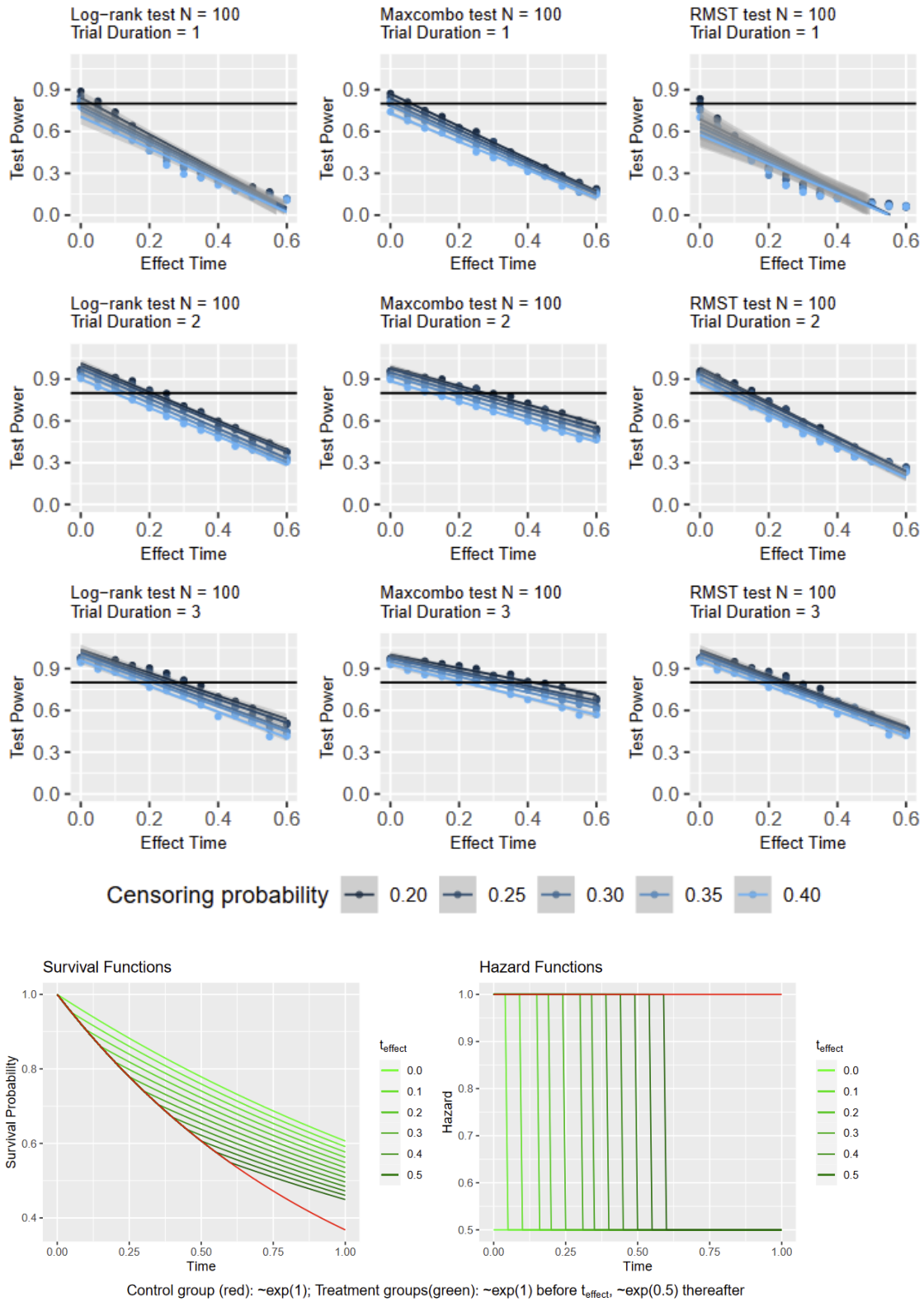
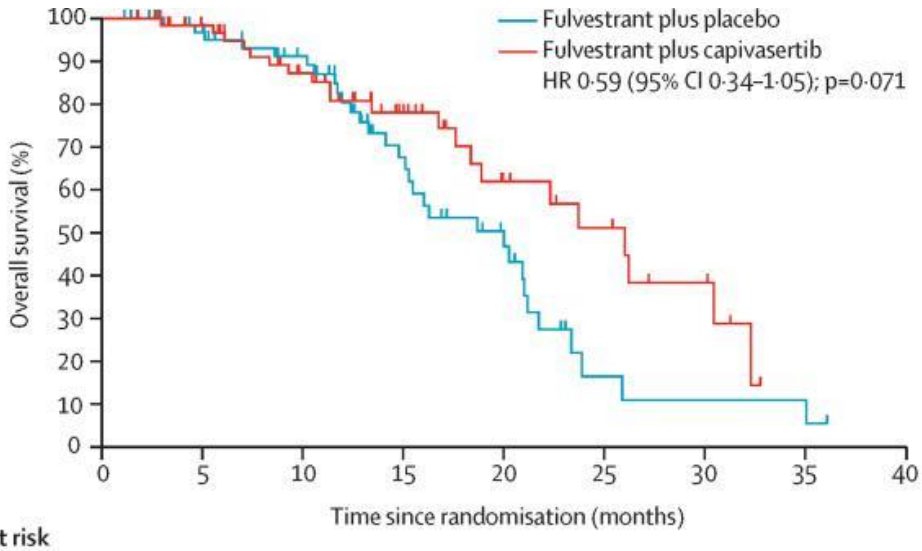
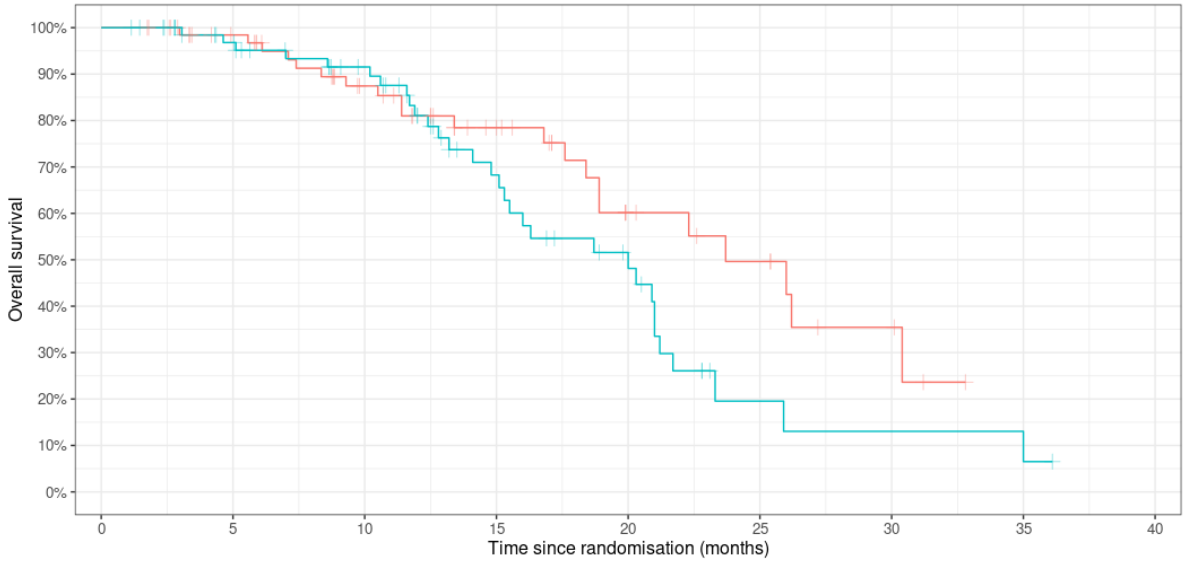


Figure 5. Kaplan–Meier curves of overall survival Jones et al.




| Number at risk (number censored) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Fulvestrant plus placebo | 71 (0) | 57 (12) | 44 (22) | 24 (32) | 14 (36) | 3 (39) | 2 (39) | 2 (39) | 0 (40) |
| Fulvestrant plus capivasertib | 69 (0) | 56 (12) | 42 (20) | 24 (34) | 13 (41) | 9 (43) | 5 (45) | 0 (48) | 0 (48) |

Figure 6. Reconstructed Kaplan–Meier curves of overall survival Jones et al.




| At Risk (Censored) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Fulvestrant plus capivasertib | 69 (0) | 58 (10) | 42 (20) | 27 (32) | 13 (40) | 9 (42) | 4 (45) | 0 (48) | 0 (48) |
| Fulvestrant plus placebo | 71 (0) | 59 (10) | 46 (20) | 25 (31) | 15 (35) | 3 (39) | 2 (39) | 2 (39) | 0 (40) |

# TABLES

Table 1. Power of Log-Rank vs MC vs RMST for Crossing Curves

| Test Power (Crossing Curves) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | N = 50 | | | N = 100 | | | N = 200 | |
| $P_{censoring}$ | Shape | Logrank | MaxCombo | RMST | Logrank | MaxCombo | RMST | Logrank | MaxCombo | RMST |
| Trial Duration: 1 | | | | | | | | | |
| 0.2 | 1 | 0.197 | 0.183 | 0.195 | 0.351 | 0.328 | 0.323 | 0.593 | 0.576 | 0.547 |
| 0.2 | 2 | 0.709 | 0.721 | 0.896 | 0.951 | 0.967 | 0.995 | 0.999 | 1.000 | 1.000 |
| 0.2 | 3 | 0.942 | 0.958 | 0.997 | 0.999 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.2 | 4 | 0.993 | 0.994 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.4 | 1 | 0.182 | 0.159 | 0.171 | 0.279 | 0.261 | 0.221 | 0.487 | 0.475 | 0.432 |
| 0.4 | 2 | 0.611 | 0.596 | 0.801 | 0.880 | 0.905 | 0.980 | 0.992 | 0.996 | 1.000 |
| 0.4 | 3 | 0.863 | 0.864 | 0.964 | 0.993 | 0.995 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.4 | 4 | 0.954 | 0.954 | 0.996 | 0.999 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 |
| Trial Duration: 2 | | | | | | | | | |
| 0.2 | 1 | 0.288 | 0.262 | 0.283 | 0.494 | 0.463 | 0.491 | 0.781 | 0.759 | 0.741 |
| 0.2 | 2 | 0.288 | 0.533 | 0.595 | 0.505 | 0.880 | 0.879 | 0.800 | 0.997 | 0.998 |
| 0.2 | 3 | 0.227 | 0.836 | 0.764 | 0.350 | 0.997 | 0.961 | 0.599 | 1.000 | 0.999 |
| 0.2 | 4 | 0.162 | 0.976 | 0.810 | 0.228 | 1.000 | 0.977 | 0.410 | 1.000 | 1.000 |
| 0.4 | 1 | 0.238 | 0.214 | 0.245 | 0.391 | 0.371 | 0.383 | 0.649 | 0.622 | 0.623 |
| 0.4 | 2 | 0.199 | 0.343 | 0.389 | 0.399 | 0.670 | 0.676 | 0.668 | 0.961 | 0.952 |
| 0.4 | 3 | 0.178 | 0.609 | 0.532 | 0.281 | 0.940 | 0.812 | 0.523 | 1.000 | 0.981 |
| 0.4 | 4 | 0.138 | 0.813 | 0.545 | 0.201 | 0.997 | 0.842 | 0.307 | 1.000 | 0.985 |
| Trial Duration: 3 | | | | | | | | | |
| 0.2 | 1 | 0.269 | 0.281 | 0.270 | 0.531 | 0.489 | 0.529 | 0.806 | 0.787 | 0.800 |
| 0.2 | 2 | 0.170 | 0.557 | 0.284 | 0.264 | 0.918 | 0.528 | 0.464 | 0.999 | 0.776 |
| 0.2 | 3 | 0.126 | 0.894 | 0.324 | 0.188 | 0.999 | 0.541 | 0.331 | 1.000 | 0.789 |
| 0.2 | 4 | 0.143 | 0.980 | 0.421 | 0.185 | 1.000 | 0.594 | 0.311 | 1.000 | 0.812 |
| 0.4 | 1 | 0.236 | 0.223 | 0.249 | 0.402 | 0.382 | 0.410 | 0.731 | 0.706 | 0.718 |
| 0.4 | 2 | 0.118 | 0.401 | 0.143 | 0.235 | 0.744 | 0.259 | 0.372 | 0.964 | 0.420 |
| 0.4 | 3 | 0.120 | 0.699 | 0.166 | 0.171 | 0.979 | 0.243 | 0.269 | 1.000 | 0.321 |
| 0.4 | 4 | 0.117 | 0.868 | 0.260 | 0.170 | 0.999 | 0.329 | 0.260 | 1.000 | 0.410 |

Table 2. Power of Log-Rank vs MC vs RMST for Early Effect

| Test Power (Early Effect) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | N = 50 | | | N = 100 | | | N = 200 | | |
| $P_{censoring}$ | $dur_{effect}$ | Logrank | MaxCombo | RMST | Logrank | MaxCombo | RMST | Logrank | MaxCombo | RMST |
| Trial Duration: 1 | | | | | | | | | | |
| 0.2 | 0.3 | 0.161 | 0.174 | 0.254 | 0.270 | 0.306 | 0.450 | 0.454 | 0.537 | 0.730 |
| 0.2 | 0.5 | 0.328 | 0.332 | 0.485 | 0.562 | 0.595 | 0.762 | 0.852 | 0.906 | 0.971 |
| 0.2 | 0.7 | 0.542 | 0.524 | 0.671 | 0.822 | 0.817 | 0.918 | 0.984 | 0.983 | 0.994 |
| 0.2 | 0.9 | 0.732 | 0.698 | 0.737 | 0.954 | 0.949 | 0.952 | 0.998 | 0.998 | 0.998 |
| 0.4 | 0.3 | 0.151 | 0.142 | 0.213 | 0.216 | 0.217 | 0.346 | 0.369 | 0.399 | 0.595 |
| 0.4 | 0.5 | 0.266 | 0.259 | 0.391 | 0.478 | 0.469 | 0.652 | 0.744 | 0.774 | 0.918 |
| 0.4 | 0.7 | 0.416 | 0.390 | 0.540 | 0.711 | 0.691 | 0.821 | 0.954 | 0.957 | 0.990 |
| 0.4 | 0.9 | 0.590 | 0.557 | 0.596 | 0.882 | 0.866 | 0.882 | 0.991 | 0.990 | 0.990 |
| Trial Duration: 2 | | | | | | | | | | |
| 0.2 | 0.3 | 0.123 | 0.153 | 0.148 | 0.213 | 0.262 | 0.262 | 0.342 | 0.474 | 0.436 |
| 0.2 | 0.5 | 0.235 | 0.287 | 0.321 | 0.445 | 0.550 | 0.574 | 0.724 | 0.848 | 0.846 |
| 0.2 | 0.7 | 0.383 | 0.440 | 0.505 | 0.645 | 0.765 | 0.793 | 0.915 | 0.976 | 0.980 |
| 0.2 | 0.9 | 0.493 | 0.537 | 0.634 | 0.795 | 0.864 | 0.910 | 0.987 | 0.998 | 0.999 |
| 0.4 | 0.3 | 0.091 | 0.108 | 0.120 | 0.174 | 0.193 | 0.212 | 0.293 | 0.363 | 0.357 |
| 0.4 | 0.5 | 0.175 | 0.197 | 0.224 | 0.342 | 0.389 | 0.424 | 0.586 | 0.684 | 0.695 |
| 0.4 | 0.7 | 0.314 | 0.334 | 0.385 | 0.530 | 0.605 | 0.659 | 0.822 | 0.893 | 0.906 |
| 0.4 | 0.9 | 0.404 | 0.427 | 0.505 | 0.700 | 0.741 | 0.805 | 0.932 | 0.964 | 0.975 |
| Trial Duration: 3 | | | | | | | | | | |
| 0.2 | 0.3 | 0.131 | 0.173 | 0.142 | 0.173 | 0.265 | 0.187 | 0.321 | 0.501 | 0.354 |
| 0.2 | 0.5 | 0.231 | 0.275 | 0.253 | 0.387 | 0.528 | 0.437 | 0.667 | 0.852 | 0.733 |
| 0.2 | 0.7 | 0.370 | 0.445 | 0.430 | 0.626 | 0.773 | 0.697 | 0.868 | 0.969 | 0.927 |
| 0.2 | 0.9 | 0.477 | 0.547 | 0.541 | 0.751 | 0.871 | 0.840 | 0.968 | 0.997 | 0.985 |
| 0.4 | 0.3 | 0.089 | 0.089 | 0.085 | 0.177 | 0.196 | 0.166 | 0.231 | 0.324 | 0.208 |
| 0.4 | 0.5 | 0.154 | 0.183 | 0.156 | 0.305 | 0.377 | 0.296 | 0.536 | 0.687 | 0.515 |
| 0.4 | 0.7 | 0.264 | 0.294 | 0.282 | 0.469 | 0.545 | 0.461 | 0.749 | 0.876 | 0.755 |
| 0.4 | 0.9 | 0.400 | 0.430 | 0.410 | 0.661 | 0.763 | 0.688 | 0.895 | 0.947 | 0.906 |

Table 3. Power of Log-Rank vs MC vs RMST for Late Effect

| Test Power (Late Effect) | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | N = 50 | | | N = 100 | | | N = 200 | | |
| $P_{censoring}$ | $t_{effect}$ | Logrank | MaxCombo | RMST | Logrank | MaxCombo | RMST | Logrank | MaxCombo | RMST |
| Trial Duration: 1 | | | | | | | | | | |
| 0.2 | 0.0 | 0.582 | 0.560 | 0.529 | 0.889 | 0.873 | 0.835 | 0.992 | 0.989 | 0.978 |
| 0.2 | 0.2 | 0.315 | 0.367 | 0.213 | 0.546 | 0.628 | 0.350 | 0.833 | 0.905 | 0.603 |
| 0.2 | 0.4 | 0.157 | 0.212 | 0.085 | 0.274 | 0.376 | 0.129 | 0.494 | 0.696 | 0.211 |
| 0.2 | 0.6 | 0.081 | 0.098 | 0.067 | 0.116 | 0.189 | 0.059 | 0.189 | 0.306 | 0.074 |
| 0.4 | 0.0 | 0.469 | 0.430 | 0.415 | 0.778 | 0.742 | 0.702 | 0.969 | 0.961 | 0.940 |
| 0.4 | 0.2 | 0.269 | 0.318 | 0.195 | 0.464 | 0.543 | 0.321 | 0.738 | 0.823 | 0.498 |
| 0.4 | 0.4 | 0.129 | 0.174 | 0.083 | 0.216 | 0.313 | 0.117 | 0.407 | 0.582 | 0.184 |
| 0.4 | 0.6 | 0.061 | 0.082 | 0.048 | 0.106 | 0.148 | 0.059 | 0.167 | 0.255 | 0.081 |
| Trial Duration: 2 | | | | | | | | | | |
| 0.2 | 0.0 | 0.746 | 0.719 | 0.731 | 0.967 | 0.959 | 0.958 | 1.000 | 1.000 | 0.999 |
| 0.2 | 0.2 | 0.537 | 0.582 | 0.453 | 0.822 | 0.849 | 0.738 | 0.982 | 0.990 | 0.952 |
| 0.2 | 0.4 | 0.324 | 0.403 | 0.278 | 0.599 | 0.728 | 0.457 | 0.870 | 0.956 | 0.745 |
| 0.2 | 0.6 | 0.224 | 0.302 | 0.158 | 0.377 | 0.544 | 0.254 | 0.648 | 0.852 | 0.462 |
| 0.4 | 0.0 | 0.642 | 0.623 | 0.616 | 0.905 | 0.897 | 0.894 | 0.993 | 0.989 | 0.990 |
| 0.4 | 0.2 | 0.434 | 0.453 | 0.399 | 0.694 | 0.738 | 0.616 | 0.929 | 0.962 | 0.888 |
| 0.4 | 0.4 | 0.265 | 0.322 | 0.229 | 0.479 | 0.597 | 0.401 | 0.788 | 0.891 | 0.670 |
| 0.4 | 0.6 | 0.203 | 0.271 | 0.177 | 0.306 | 0.465 | 0.234 | 0.539 | 0.766 | 0.399 |
| Trial Duration: 3 | | | | | | | | | | |
| 0.2 | 0.0 | 0.817 | 0.802 | 0.812 | 0.978 | 0.974 | 0.975 | 1.000 | 1.000 | 1.000 |
| 0.2 | 0.2 | 0.627 | 0.642 | 0.612 | 0.904 | 0.921 | 0.880 | 0.997 | 0.999 | 0.995 |
| 0.2 | 0.4 | 0.405 | 0.527 | 0.400 | 0.698 | 0.808 | 0.659 | 0.945 | 0.988 | 0.918 |
| 0.2 | 0.6 | 0.275 | 0.397 | 0.261 | 0.509 | 0.684 | 0.457 | 0.814 | 0.940 | 0.773 |
| 0.4 | 0.0 | 0.699 | 0.689 | 0.710 | 0.943 | 0.928 | 0.943 | 0.996 | 0.995 | 0.996 |
| 0.4 | 0.2 | 0.477 | 0.517 | 0.504 | 0.766 | 0.800 | 0.765 | 0.977 | 0.984 | 0.971 |
| 0.4 | 0.4 | 0.339 | 0.429 | 0.366 | 0.556 | 0.677 | 0.574 | 0.895 | 0.946 | 0.890 |
| 0.4 | 0.6 | 0.233 | 0.319 | 0.249 | 0.415 | 0.569 | 0.419 | 0.692 | 0.871 | 0.691 |

Table 4. Type I error

| Type I error | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | N = 50 | | | N = 100 | | | N = 200 | | |
| $P_{censoring}$ | Logrank | MaxCombo | RMST | Logrank | MaxCombo | RMST | Logrank | MaxCombo | RMST |
| Trial Duration: 1 | | | | | | | | | |
| 0.2 | 0.043 | 0.046 | 0.056 | 0.056 | 0.053 | 0.061 | 0.055 | 0.047 | 0.050 |
| 0.4 | 0.041 | 0.043 | 0.060 | 0.050 | 0.055 | 0.041 | 0.047 | 0.041 | 0.055 |
| Trial Duration: 2 | | | | | | | | | |
| 0.2 | 0.050 | 0.048 | 0.055 | 0.052 | 0.055 | 0.058 | 0.059 | 0.048 | 0.065 |
| 0.4 | 0.049 | 0.049 | 0.059 | 0.062 | 0.060 | 0.066 | 0.042 | 0.047 | 0.044 |
| Trial Duration: 3 | | | | | | | | | |
| 0.2 | 0.044 | 0.053 | 0.048 | 0.069 | 0.069 | 0.066 | 0.059 | 0.058 | 0.059 |
| 0.4 | 0.048 | 0.053 | 0.060 | 0.058 | 0.047 | 0.058 | 0.055 | 0.056 | 0.052 |

# REFERENCES


Castañon E, Sanchez-Arraez A, Alvarez-Manceñido F, Jimenez-Fonseca P, and Carmona-Bayonas A. Critical reappraisal of phase III trials with immune checkpoint inhibitors in non-proportional hazards settings. Eur J Cancer. 2020;136:159-168.

Dormuth, I., Liu, T., Xu, J., *et al.* Which test for crossing survival curves? A user's guideline. *BMC Med Res Methodol* **22**, 34 (2022). https://doi.org/10.1186/s12874-022-01520-0

Dormuth, I., Liu, T., Xu, J., Pauly, M., & Ditzhaus, M. (2023). A comparative study to alternatives to the log-rank test. Contemporary clinical trials, 128, 107165. https://doi.org/10.1016/j.cct.2023.107165

Duke-Center for Health Policy at Duke University (2018). Public workshop: oncology clinical trials in the presence of non-proportional hazards, https://healthpolicy.duke.edu/sites/default/files/2020-03/oncology_trials_workshop_discussion_guide_final.pdf

Eaton A, Therneau T, and Le-Rademacher J. Designing clinical trials with a (restricted) mean survival time endpoint: Practical considerations. Clinical Trials. 2020;17(3):285-294. doi:10.1177/1740774520905563.

Flandre P. and O'Quigley J. Comparing Kaplan–Meier curves with delayed treatment effects: applications in immunotherapy trials. Appl Statist 2019; 68:915–939.

Guyot, P., Ades, A. E., Ouwens, M. J., & Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. BMC Medical Research Methodology, 12, 9. https://doi.org/10.1186/1471-2288-12-9.

Han K., Jung I. Restricted Mean Survival Time for Survival Analysis: A Quick Guide for Clinical Researchers. Korean J Radiol. 2022 May;23(5):495-499. https://doi.org/10.3348/kjr.2022.0061.

Harhay MO, Porcher R, Cantu E, Crowther MJ, Christie JD, Thabut G, et al. An alternative approach for the analysis of time-to-event and survival outcomes in pulmonary medicine. Am J Respir Crit Care Med 2018;198:684-687.

Huang B and Kuan PF. Comparison of the restricted mean survival time with the hazard ratio in superiority trials with a time-to-event endpoint. Pharm Stat 2018; 17(3): 202–213.

Huang, B., Wei, L. J., & Ludmir, E. B. (2020). Estimating Treatment Effect as the Primary Analysis in a Comparative Study: Moving Beyond P Value. Journal of clinical oncology: official journal of the American Society of Clinical Oncology, 38(17), 2001–2002. https://doi.org/10.1200/JCO.19.03111

Jiang, Z., Wang, L., Li, C., Xia, J., & Jia, H. (2012). A practical simulation method to calculate sample size of group sequential trials for time-to-event data under exponential and Weibull distribution. PloS one, 7(9), e44013. https://doi.org/10.1371/journal.pone.0044013.

Jones, R. H., Casbard, A., Carucci, M., Cox, C., Butler, R., Alchami, F., Madden, T. A., Bale, C., Bezecny, P., Joffe, J., Moon, S., Twelves, C., Venkitaraman, R., Waters, S., Foxley, A., & Howell, S. J. (2020). Fulvestrant plus capivasertib versus placebo after relapse or progression on an aromatase inhibitor in metastatic, oestrogen receptor-positive breast cancer (FAKTION): a multicentre, randomised, controlled, phase 2 trial. The Lancet. Oncology, 21(3), 345–357. https://doi.org/10.1016/S1470-2045(19)30817-4

Kim DH, Uno H, Wei LJ. Restricted mean survival time as a measure to interpret clinical trial results. JAMA Cardiol 2017;2:1179-1180.

Kristiansen, I. PRM39 survival curve convergences and crossing: a threat to validity of meta-analysis? Value Health 15, A652 (2012).

Lin, Ray S., Ji Lin, Satrajit Roychoudhury, Keaven M. Anderson, Tianle Hu, Bo Huang, Larry F. Leon, et al. 2020. "Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards: A Comparative Analysis." Statistics in Biopharmaceutical Research 12 (2): 187–98.

Li H, Han D, Hou Y, Chen H, and Chen Z. Statistical inference methods for two crossing survival curves: a comparison of methods. PLoS One 2015; 10: e0116774.

Mukhopadhyay P, Ye J, Anderson KM, Roychoudhury S, Rubin EH, Halabi S, Chappell RJ. Log-Rank Test vs. MaxCombo and Difference in Restricted Mean Survival Time Tests for Comparing Survival Under Nonproportional Hazards in Immuno-oncology Trials: A Systematic Review and Meta-analysis. JAMA Oncol. 2022 Sep 1;8(9):1294-1300. doi: 10.1001/jamaoncol.2022.2666. PMID: 35862037; PMCID: PMC9305601.

Pak K., Uno H, Kim DH, Tian L, Kane RC, Takeuchi M, et al. Interpretability of cancer clinical trial results using restricted mean survival time as an alternative to the hazard ratio. JAMA Oncol 2017;3:1692-1696.

Perego C, Sbolli M, Specchia C, Fiuzat M, McCaw ZR, Metra M, et al. Utility of restricted mean survival time analysis for heart failure clinical trial evaluation and interpretation. JACC Heart Fail 2020;8:973-983.

Qiu P, Sheng J. A two-stage procedure for comparing hazard rate functions. J R Stat Soc Series B Stat Methodol 2008; 70: 191-208. 20.

Roychoudhury S, Anderson KM, Ye J, and Mukhopadhyay P. Robust design and analysis of clinical trials with nonproportional hazards: a straw man guidance from a cross-pharma working group. Stat Biopharm Res. 2021;1-15. doi:10.1080/19466315.2021.1874507

Royston, P., & Parmar, M. K. (2020). A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome. Trials, 21(1), 315. https://doi.org/10.1186/s13063-020-4153-2

Royston P, Parmar MK. 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 Med Res Methodol 2013;13:152.

Tian, L., Fu, H., Ruberg, S. J., Uno, H., & Wei, L. J. (2018). Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. Biometrics, 74(2), 694–702. https://doi.org/10.1111/biom.12770

Trinquart, L., Jacot, J., Conner, S. C., & Porcher, R. (2016). Comparison of Treatment Effects Measured by the Hazard Ratio and by the Ratio of Restricted Mean Survival Times in Oncology Randomized Controlled Trials. Journal of clinical oncology: official journal of the American Society of Clinical Oncology, 34(15), 1813–1819. https://doi.org/10.1200/JCO.2015.64.2488

Uno H, Claggett B, Tian L, Inoue E, Gallo P, Miyata T, et al. Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J Clin Oncol 2014;32:2380-2385.