Robust Group Sequential Designs for Trials with Survival Endpoints and Delayed Response

Pranab Ghosh[1], Robin Ristl[2], Franz König[2], Martin Posch[2], Christopher Jennison[3], Heiko Götte[4], Armin Schüler[4], Cyrus Mehta[5,6*]

[1]Pfizer Corporation, Cambridge MA, USA

[2] Medical University of Vienna, Austria

[3] University of Bath, UK

[4] Merck AG, Darmstadt, Germany

[5]Cytel Inc, Cambridge MA, USA

[6]Harvard TH Chan School of Public Health, Boston MA, USA[6]

May 30, 2020

Abstract

Randomized clinical trials in oncology typically utilize time-to-event endpoints such as progression-free survival (PFS) or overall survival (OS) as their primary efficacy endpoints, and the most commonly used statistical test to analyze these endpoints is the logrank test. The power of the logrank test depends on the behaviour of the hazard ratio of the treatment arm to the control arm. Under the assumption of proportional hazards (PH) , the logrank test is asymptotically fully efficient. However, this proportionality assumption does not hold true if there is a delayed treatment effect. Cancer immunology has evolved over time and several cancer vaccines are available in the market for treating existing cancers. This includes sipuleucel-T for metastatic hormone-refractory prostate cancer, nivolumab for metastatic melanoma, and pembrolizumab for advanced non-small-cell lung cancer. Since cancer vaccines require some time to elicit an immune response, a delayed treatment effect is observed, resulting in a violation of the proportional hazards assumption. Thus the traditional logrank test may not be optimal for testing immuno-oncology drugs in randomized clinical trials. Moreover, the new immuno-oncology compounds have been shown to be very effective in prolonging overall survival. Therefore it is desirable to implement a group sequential design with the possibility of early stopping for overwhelming efficacy. In this paper we investigate the max-combo test, which utilzes the maximum of three weighted log-rank statistics, as a robust alternative to the logrank test. The new test is implemented for two-stage designs with possible early stopping at the interim analysis time point.

Key Words: immuno-oncology, non-proportional hazards, delayed response, late separation of survival curves, log-rank test, Harrington-Fleming test, max-combo test, group sequential design

*Corresponding Author; mehta@cytel.com

1 Introduction

Consider a clinical trial where a new treatment will be compared to the standard of care, and where the primary endpoint is overall survival time (OS) or progression free survival time (PFS). Let $h_0(t)$ and $h_1(t)$ represent the hazard rates at any time t for the standard of care and the new treatment, respectively, and let $\lambda(t) = \frac{h_1(t)}{h_0(t)}$ represent the hazard ratio at time t. We wish to test the null hypothesis H_0: $h_1(t) = h_0(t)$ for all t, against the 1-sided alternative hypothesis $h_1(t) \leq h_0(t)$ with strict inequality at at least one value of t. Under the proportional hazard assumption, the hazard ratio $\lambda(t) = \lambda$, independent of t, and the alternative hypothesis implies that $\lambda < 1$. In this setting the logrank test is asymptotically fully efficient. (See for example Kalbfleisch and Prentice [2011]). Sometimes, however, the treatment effect takes time to materialize, which results into a time lag before the two survival curves separate. For example, the immunology drugs being tested in oncology trials exhibit this phenomenon.This late separation or delayed treatment effect implies that the assumption of proportional hazards no longer holds. Another example of non-proportional hazards is when the hazard rates for survival change after disease progression.

When the proportional hazards assumtion is violated, the logrank test is no longer fully efficient and may lose power relative to other tests that are better able to handle the non-proportionality. The Fleming-Harrington or $G^{\rho,\gamma}$ class of hypothesis tests (Fleming and Harrington [1981]) are generalizations of the logrank test in which weights are assigned to the failure times by choice of two parameters, ρ and γ. Thus they are known as weighted logrank tests. The $G^{0,0}$ test applies equal weights to the failure times and thereby yields the standard logrank test. The $G^{1,0}$ and $G^{0,1}$ tests are more sensitive, respectively, to early and late difference alternatives. Since one would not know in advance whether the survival curves will separate early, separate late, or exhibit the proportional hazards alternative, a robust option for hypothesis testing when non-proportional hazards are expected, is to define the test statistic as the maximum of the above three Fleming-Harrington statistics. This test is referred to as the max-combo test. It was proposed in an Industry and FDA Sponsored Public Workshop in Washington DC in 2017duk. Several presentations at that workshop compared relative efficiencies of the max-combo and logrank tests for non-proportional hazards alternatives. These comparisons, however, were restricted to single-look, fixed-sample designs. Several recent clinical trials of immuno-oncology compounts, for example the PD-1 inhibitors (see McDermott and Jimeno [2015]), have been shown to prolong survival significantly relative conventional cytotoxic therapies. Therefore the option to perform an interim analysis and stop early if there is overwhelming efficacy would be desirable. In this paper we will compare the performance of the max-combo test to that of the logrank test for group sequential designs with the possibility of early efficacy stopping at an interim analysis time point.

2 Modelling the Survival Curves

Ristl et al (2019) have developed non-proportional hazards models by specifying different hazard rates for the treatment and control arms before and after disease progression, classified by biomarker status, as shown in Figure 1. For a delayed onset of treatment effect after a time t_{onset}, the hazard function is

$$\lambda(t) = \lambda_{pre-onset} I_{t<t_{onset}} + \lambda_{post-onset} I_{t\geq onset},$$

where $I_{(.)}$ is the indicator function. To model changing hazards after disease progression, let Y denote the time to disease progression based on the hazard functions $\lambda_p(t) \in \{\lambda_p, \lambda_{p-}(t), \lambda_{p+}(t)\}$. Let $\lambda_D(t) \in \{\lambda_D, \lambda_{D-}(t), \lambda_{D+}(t)\}$ be the hazard functions for death before disease progression. Let $\lambda_{PD}(t) \in \{\lambda_{PD}, \lambda_{PD-}(t), \lambda_{PD+}(t)\}$ be the hazard functions for death after disease progression. Conditional on $Y = s$ the hazard function for death is

$$\lambda(t|Y = s) = \lambda_D(t) I_{t\leq s} + \lambda_{PD}(t) I_{t>s}$$

and the corresponding conditional survival function is

$$S(t|Y = s) = \exp\left\{-\int_0^t \lambda(t|Y = s)\right\} ds \ .$$

Thus the unconditional survival function is

$$S(t) = \int_0^t S(t|Y = s) dP(Y = s) \ .$$

The survival distribution comprising biomarker negative and positive subpopulations is obtained as a mixture distribution over the survival functions of the subpopulations. To simulate data, survival times are sampled from the theoretical overall survival function $S(t)$. We shall utilize these models to generate the late separation of the survival curves under two scenarios; delayed response and changing hazards after disease progression. We shall assume that the relevant hazard rates are piece-wise constant.

2.1 Delayed Treatment Effect

For this scenario we will assume that the median survival on both arms is 12 months for the first 3 months. After 3 months, the curves separate with hazard ratio 0.6. Figure 2 displays the relevant survival, hazard, and hazard ratio functions.

2.2 Changing Hazards after Disease Progression

For this scenario we assume that the median survival time corresponding to λ_D is 18 months for the control arm. The hazard ratio is 1 for the first 3 months, dropping to 0.6 after 3

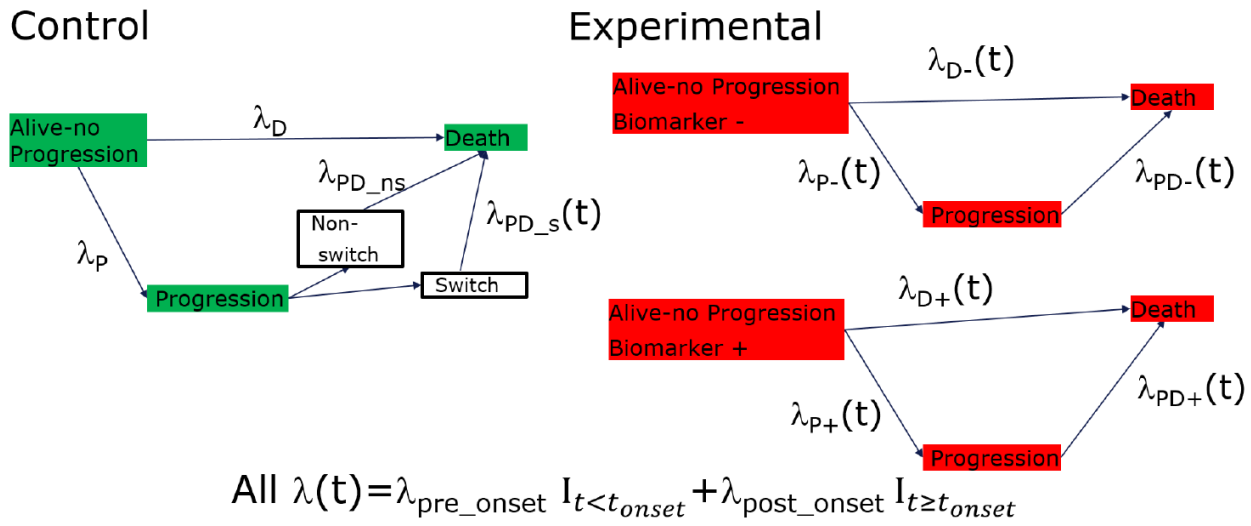


Figure 1: Multi-state representation of the modeled sources of non-proportional hazards

months. The median time to progression corresponding λ_p is 5 months for the control arm. The hazard ratio is 1 for the first 4 months, dropping to 0.4 after 3 months. The median survival time after progression corresponding to λ_{PD} is 11 months for the control arm with a hazard ratio of 0.5. The survival, hazard rate and hazard ratio functions arising from these specifications are displayed in Figure 3.

3 The Max-Combo Statistics

The statistical power of the logrank test deteriorates when the proportional hazards assumption is violated. However a weighted version of log-rank test, with properly chosen weights, will regain the lost statistical power. Suppose we are analyzing survival data at some calendar time t and have already observed $d_1, d_2, \ldots d_k$ events of interest at corresponding *patient follow-up times* $\tau_1, \tau_2, \ldots \tau_k$. Suppose $d_{11}, d_{12}, \ldots d_{1k}$ of these events are from the treatment group. Then a weighted logrank statistic is defined as

$$G^{\rho,\gamma} = \sum_{j=1}^{k} \hat{Q}(\tau_j)\,(d_{1j} - E(d_{1j})) \tag{3.1}$$

where $E(d_{1j}) = \frac{n_{1j}d_{1j}}{n_j}$ with n_{1j} being the number of patients from the treatment arm that were at risk at time τ_j and n_j being the total number of patients at risk at time τ_j. The variance of this weighted statistic is

$$Var(G^{\rho,\gamma}) = \sum \hat{Q}(\tau_j)^2 Var(d_{1j}) = \sum \hat{Q}(\tau_j)^2 \frac{n_{1j}(n_j - n_{1j})d_j(n_j - d_j)}{n_j^2(n_j - 1)}\ . \tag{3.2}$$

In these equations, $\hat{Q}(\tau_j)$ is the weight associated with the event at time τ_j. Although, one can use any choice of weights for testing purposes, we will concentrate on the weights, proposed by

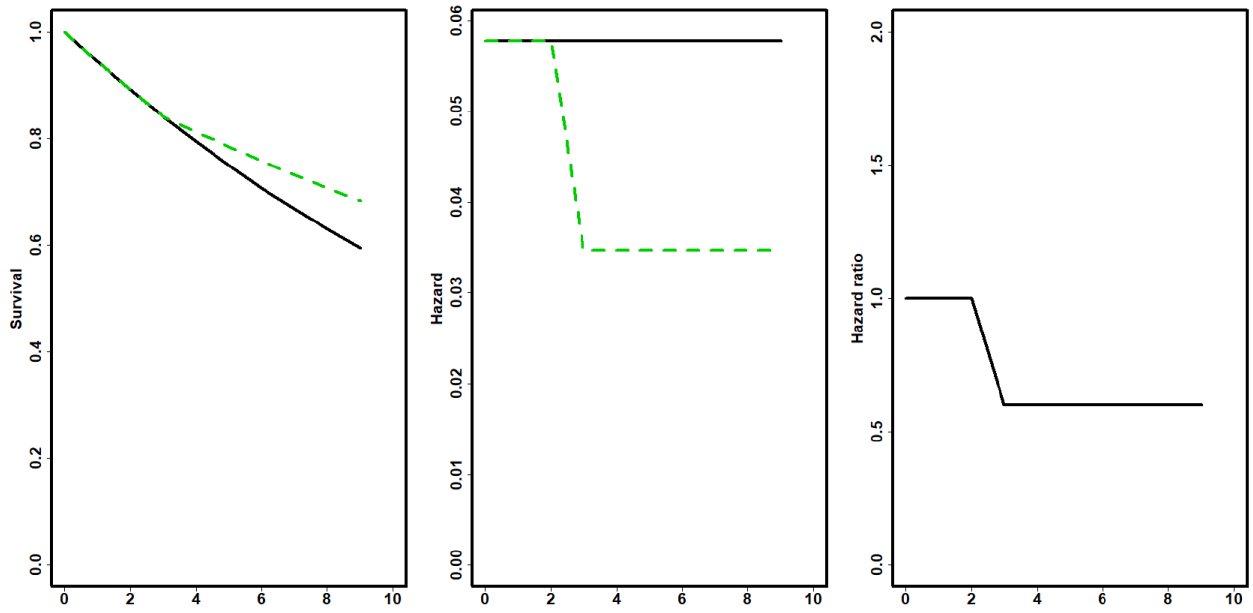


Figure 2: Survival, hazard and hazard ratio functions for late separation model

Harrington and Fleming. The Harrington-Fleming weights are parameterized by ρ and γ and can be computed as

$$\hat{Q}(\tau_j) = \hat{S}(\tau_j)^\rho \left(1 - \hat{S}(\tau_j)\right)^\gamma \tag{3.3}$$

where $\hat{S}(\tau_j)$ is the empirical survival function estimated from the pooled sample at time τ_j. The choice of ρ and γ is rather subjective and will weight certain events more heavily than others. For example, the regular log-rank test that weights each event equally uses $\rho = \gamma = 0$, whereas, $\rho = 0, \gamma = 1$ places heavier weight on the late events and almost no weight on the early events. If early events are of greater interest one would use $\rho = 1$ and $\gamma = 0$. In reality it is difficult to determine at the start of the trial whether to emphasize early or late events and hence it difficult to pre-specify a single choice of ρ and γ for the Harrington-Fleming test. An alternate testing strategy is to use several Harrington-Fleming statistics with different values of ρ and γ and take their maximum. This is the max-combo test, that was discussed in the 2017 Industry and FDA sponsored Public Workshop on Oncology Clinical Trials in the Presence of Non-Proportional Hazards. To this end we first standardized weighted log-rank statistic as

$$Z^{\rho,\gamma} = \frac{G^{\rho,\gamma}}{\sqrt{Var(G^{\rho,\gamma})}} \ .$$

The max-combo statistic is the maximum of three Harrington Fleming standardized weighted statistics:

$$\bar{Z} = \max\left(Z^{0,0}, Z^{1,0}, Z^{0,1}\right) \ . \tag{3.4}$$

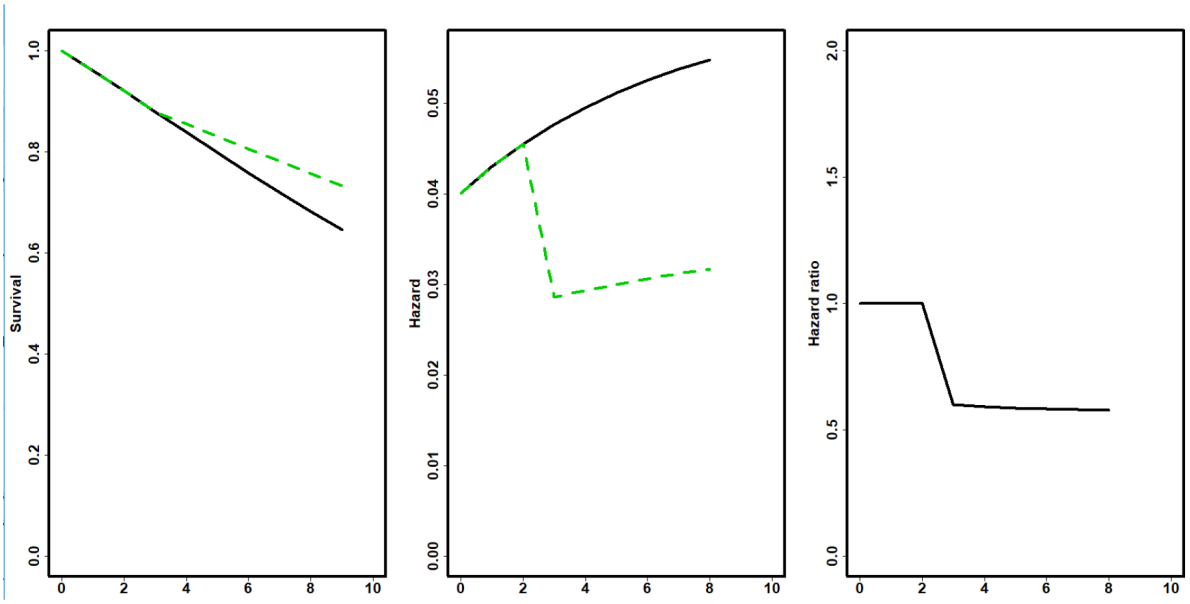


Figure 3: Survival, hazard and hazard ratio functions for disease progression model

4 Group Sequential Monitoring

Trials of immunology compounds have prolonged study durations both because the induction of immunity often takes some time and because the therapies are effective in prolonging survival. Thus a group sequential design with possible early stopping for overwhelming efficacy could be more efficient than a single-look design. Suppose interim analyses are taken at information times $t_1, t_2, \ldots\ldots, t_k$. The max-combo statistic $(\bar{Z}_k)$ at information time t_k is the maximum of $Z_k^{0,0}, Z_k^{1,0}, Z_k^{0,1}$. Although the statistical power (evaluated under the alternative hypothesis) with the max-combo test will be estimated by simulation, we must know the distribution of $\bar{Z}_k = \left(Z_k^{0,0}, Z_k^{1,0}, Z_k^{0,1}\right)$ under the null hypothesis is true to compute the group sequential efficacy boundaries. The following is the main result of this paper:

The sequentially computed max-combo statistics have the following distribution under the null hypothesis that the hazard ratio is 1:

$$E(\bar{Z}_k) = 0 \tag{4.5}$$

$$Var(Z_k^{\rho_1,\gamma_1}) = 1 \tag{4.6}$$

$$Cov(Z_k^{\rho_1,\gamma_1}, Z_k^{\rho_2,\gamma_2}) = \frac{Cov(G_k^{\rho_1,\gamma_1}, G_k^{\rho_2,\gamma_2})}{\sqrt{Var(G_k^{\rho_1,\gamma_1})Var(G_k^{\rho_2,\gamma_2})}} = \frac{Var(G_k^{(\rho_1+\rho_2)/2,(\gamma_1+\gamma_2)/2})}{\sqrt{Var(G_k^{\rho_1,\gamma_1})Var(G_k^{\rho_2,\gamma_2})}} \tag{4.7}$$

$$Cov(Z_k^{\rho_1,\gamma_1}, Z_{k+1}^{\rho_2,\gamma_2}) = \frac{Cov(G_k^{\rho_1,\gamma_1}, G_{k+1}^{\rho_2,\gamma_2})}{\sqrt{Var(G_k^{\rho_1,\gamma_1})Var(G_k^{\rho_2,\gamma_2})}} = \frac{Cov(G_k^{\rho_1,\gamma_1}, G_k^{\rho_2,\gamma_2})}{\sqrt{Var(G_k^{\rho_1,\gamma_1})Var(G_k^{\rho_2,\gamma_2})}} \tag{4.8}$$

$Var(G_k^{\rho,\gamma})$ can be computed by equation (3.2). The result for the covariance between the two

weighted statistics at the same look k, given by equation (4.7), was derived by Karrison [2016]. The result for the covariance between two weighted statistics at the two distinct looks k and $k+1$, given by equation (4.8), is based on the results in Tsiatis [1982]. This covariance structure implies that the weighted log-rank statisics have independent increments, and thus greatly facilitates the generation of group sequential efficacy boundaries.

Suppose, in a two-stage design, we decide to take the interim analysis at information time t_1 and, out of total available type I error α, we decide to spend α_1 at the interim. We can compute the early stopping boundary c_1 such that if $\bar{Z}_1 \geq c_1$ the trial will stop at time t_1 with an efficacy claim on the new treatment. To compute c_1, we need to solve the following equation

$$\begin{aligned} & P_0\left(\bar{Z}_1 \geq c_1\right) = \alpha_1 \\ \Rightarrow & P_0\left(\bar{Z}_1 < c_1\right) = P_0\left((Z_1^{0,0}, Z_1^{1,0}, Z_1^{0,1}) < c_1\right) = 1 - \alpha_1 \end{aligned} \tag{4.9}$$

Using the distribution of $(Z_1^{0,0}, Z_1^{1,0}, Z_1^{0,1})$ (equations (4.5) and (4.6)) we can compute c_1. Suppose, at the interim we fail to reject the null hypothesis due to the max-combo statistic obtained from the data being below c_1. In that case we will go to the final analysis. At the final analysis time, we need to evaluate the final stage group sequential boundary c_2 so that overall type I error is controlled at level α. Towards that end we need to solve

$$\begin{aligned} & P_0\left(\bar{Z}_1 < c_1 \cap \bar{Z}_2 \geq c_2\right) = \alpha - \alpha_1 \\ \Rightarrow & P_0\left(\bar{Z}_1 < c_1\right) - P_0\left(\bar{Z}_1 < c_1 \cap \bar{Z}_2 < c_2\right) = \alpha - \alpha_1 \\ \Rightarrow & P_0\left(\bar{Z}_1 < c_1 \cap \bar{Z}_2 < c_2\right) = 1 - \alpha \end{aligned} \tag{4.10}$$

Solving equations (4.9),(4.10) requires repeated evaluation of multi-dimensional integrals. The dimension of these integrals is three times the number of stages. The evaluation requires intensive computation and it will be difficult to solve without an efficient computation technique. We have used the approach proposed by Ghosh et al. [2017] for computing the group sequential boundaries for multi-arm multi-stage trials.

5 Simulation Experiments

We will obtain operating characteristics for the group sequential max-combo test, the group sequential log-rank test, and the single stage max-combo test by simulation under non-proportional hazards alternatives. We have modeled two different scenarios for non-proportional hazards – delayed treatment effect and changing hazards after disease progression, as explained in Section 2. For these simulation experiments we consider an immunology trial in which 300 patients are enrolled at a uniform rate of 25 patients per month over a 12 month. period. Each patient is randomized to either the new treatment or to standard of care in a 1:1 randomization ratio. After randomization, each patient is followed for up to a maximum of 30 months so that the final analysis occurs at month 42. We plan to take

one interim look at the accruing data for possible early efficacy stopping. We will examine four possible calendar times – 18 months, 21 months, 24 months and 27 months – for taking the interim look. The amount of type-1 error α_1 to be spent at the interim look is obtained from $\gamma(-5)$ spending function proposed by Hwang et al. [1990], such that

$$\alpha_1 = \alpha \frac{1 - e^{-\gamma\nu_1}}{1 - e^{-\gamma}}$$

where $\gamma = -5$ and ν_1 is the information fraction for the interim look. The information fraction ν_1 is estimated separately for each calendar time based on the enrollment rate and the hazard functions for the two treatments under the alternative hypothesis. The critical cut-off value c_2 for the final analysis is obtained by solving equation (4.10) so as to ensure that the type-1 error is α. All results are provided for $\alpha = 0.025$, one-sided.

Figure 4 compares the power of the group sequential logrank and the group sequential max-combo tests, based on the delayed response model of Section 2.1, at the four calendar times 18 months, 21 months, 24 months, and 27 months. The max-combo test is shown to

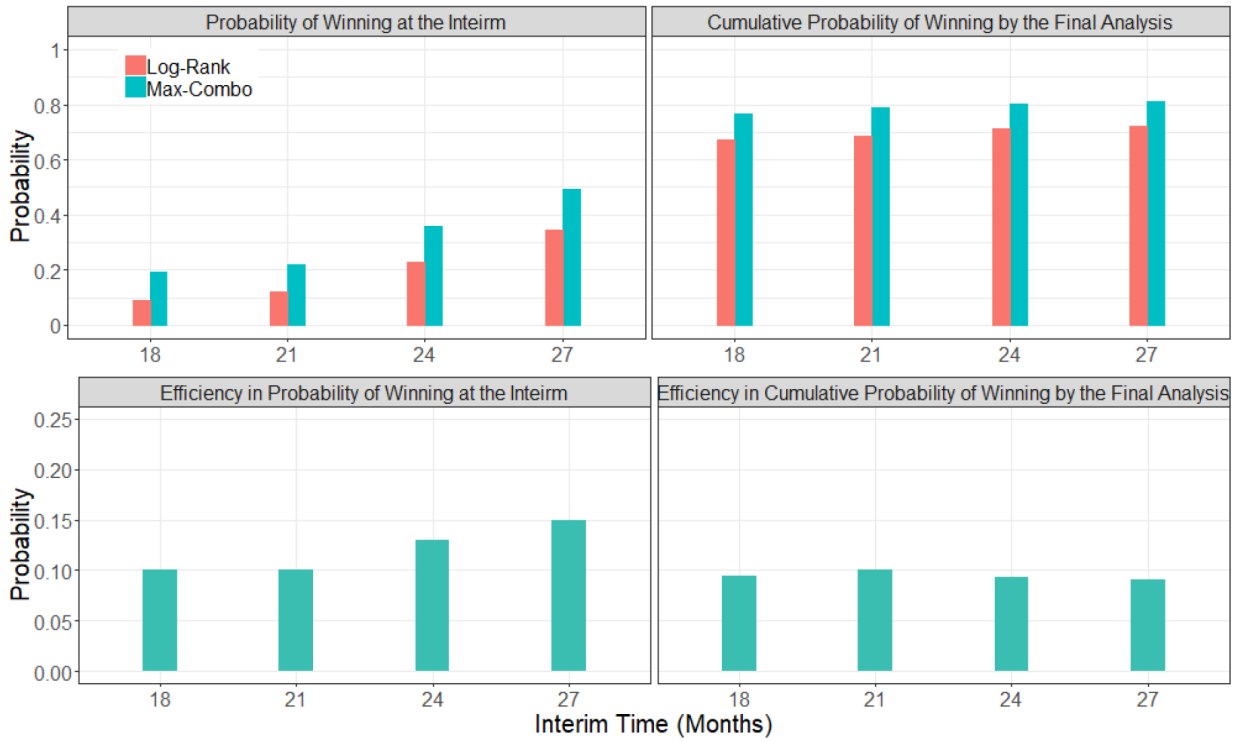


Figure 4: Power Comparisons for Delayed Response Model: logrank vs max-combo

have greater power with power gains of 10%-15% at the interim analysis and 8%-10% at the final analysis.

Figure 5 compares the power of the group sequential logrank and the group sequential max-combo tests, based on the changing hazards upon progression model of Section 2.2, at the

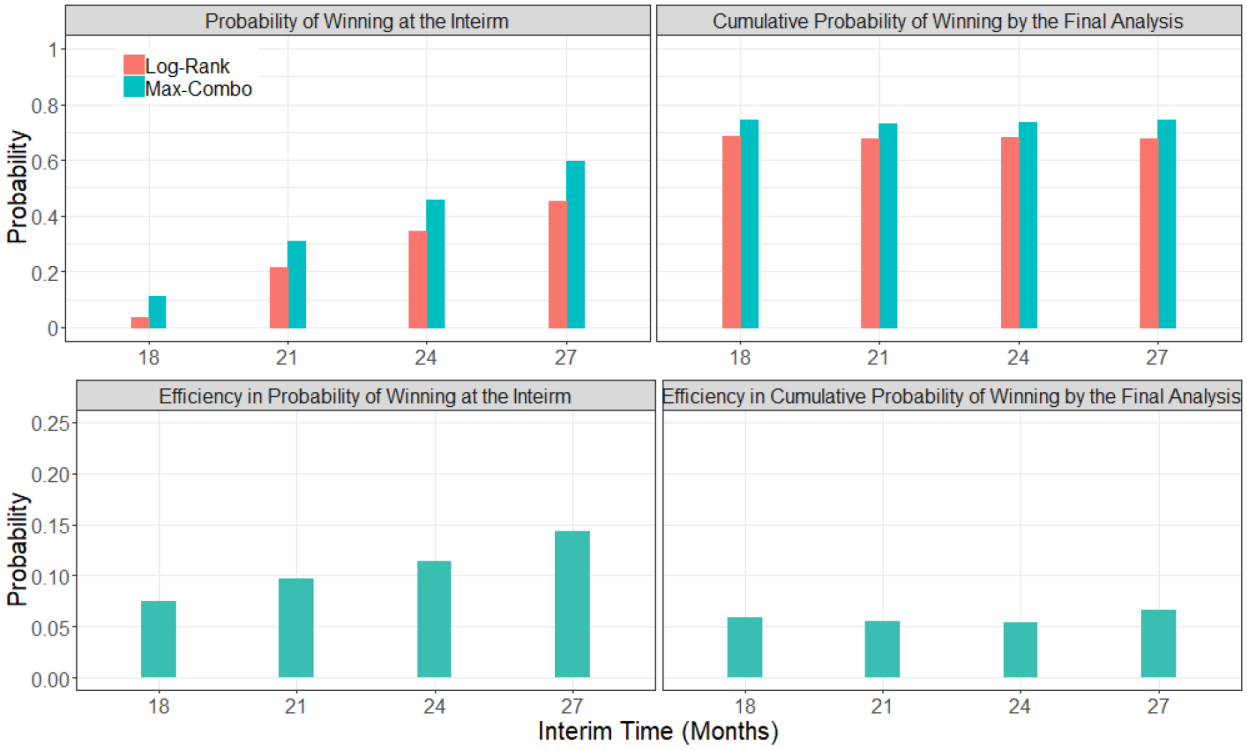


Figure 5: Power Comparisons for Changing Hazards upon Progression: logrank vs max-combo

four calendar times 18 months, 21 months, 24 months, and 27 months. The max-combo test is shown to have greater power with power gains of 5%-15% at the interim analysis and 5%-7% at the final analysis.

Since the logrank test is asymptotically fully efficient under proportional hazards, it would be interesting to know the extent to which the max-combo test loses power in this situation. This is shown in Figure 6. The max-combo test loses 3%-6% power at the interim analysis and 3% overall.

Finally, we wish to study the impact of adding an interim analysis to the max-combo test in terms of power loss for the same sample size. This is shown in Table 1 where the enrollment and follow-up strategies are the same for both designs. The overall power loss for taking the additional look is beween 1.5% and 6% for the delayed reponse model and between 1.6% and 4.5% for the changing hazards upon progression model. On the other hand there are substantial savings in sample size and study duration due to the possibility of early stopping at the interim look. For the delayed response model, there is between 19% and 49% probability of early stopping at the interim analysis time point. For the changing hazards upon progression model, the corresponding probabilities are between 12% and 59%. Thus the power losses are more than offset by the shorter study duration and smaller average sample size, and could easily be made up by committing additional patients to the study such that the average sample sizes of the two designs are the same.

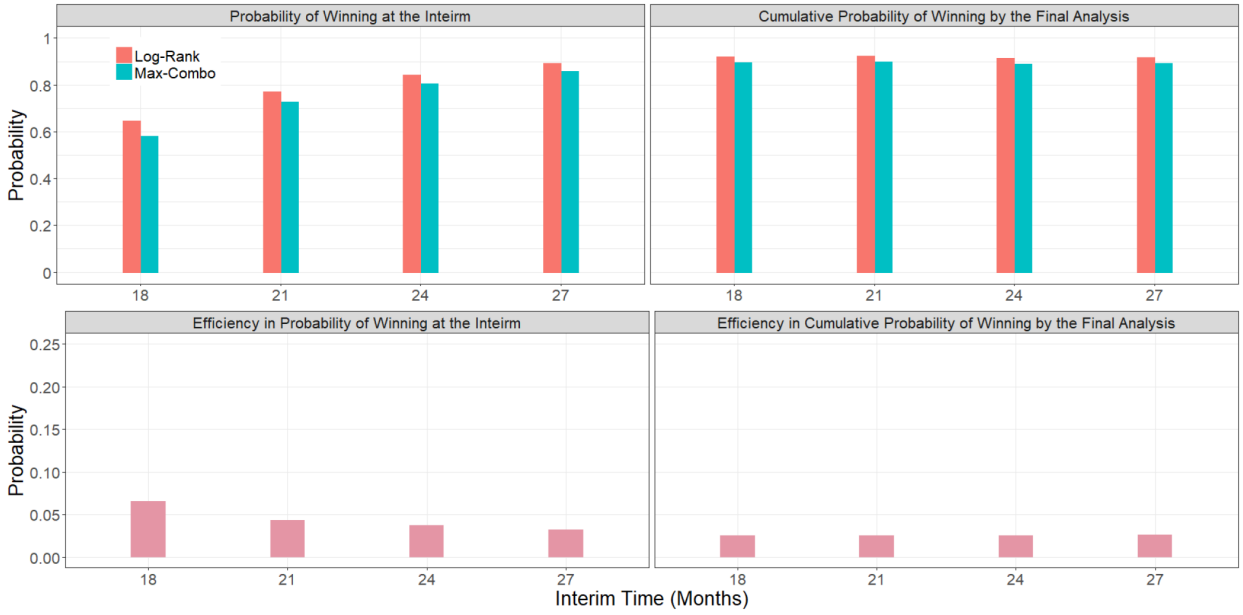


Figure 6: Power comparisons of logrank vs max-combo under proportional hazards alternatives

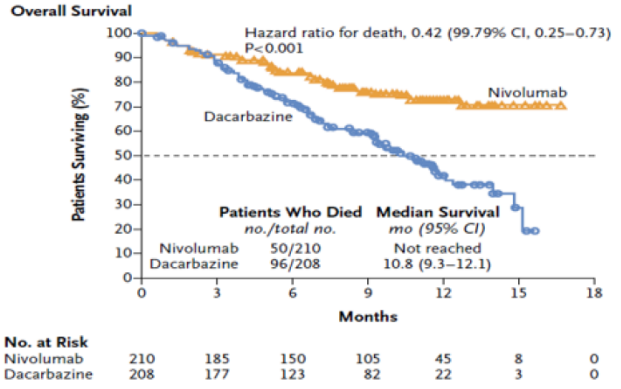
6 Discussion

Immuno-oncology trials are have met with considerable success in targeted biomarker subgroups. This very success implies that unless group sequential methods are employed for early efficacy stopping, the trial durations are likely to be considerably prolonged, resulting in delays in delivering effective new compounds to cancer patients. Group sequential methods based on the conventional logrank test might lose power, however, because of the possibility of late separation of the survival curves. Late separation is, moreover, biologically plausible either due to delayed effects of the new immunotherapies or due to changes in the hazard rates for survival following disease progression. Figure 7 displays the Kaplan-Meier curves of several recent immuno-oncology trials, and reflect both, the long study duration and the late separation of the survival curves. These results were shown at the Public Workshop for oncology clinical trials that was cited above. The max-combo test is a robust alternative to the logrank test that caters to the possibility of late separation. For two-stage group sequential designs we have shown power gains of up to 16% at the interim analysis time point and up to 8% overall for the max-combo test compared to the logrank test. As against this, if the proportional hazards assumpton holds, the max-combo test can lose up to 3% power compared to the logrank test. Finally, the saving in study duration for a two-stage design can be considerable. We have shown between 11% and 59% probabilities for early stopping under the various scenarios discussed. All these results argue for the use of tests like the max-combo test that perform better than the logrank test if the survial curves separate late while at the same time not giving away too much power if the proportional hazards alternative should hold. Finally, we have not investigated the possibility of early stopping for futility in this paper, nor

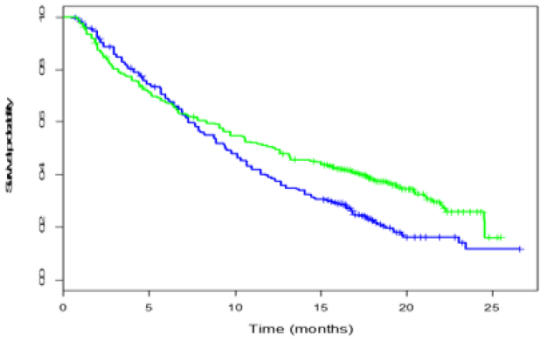
Table 1: Power of Single-Stage vs 2-Stage Max-Combo Tests

	Interim Analysis	Power		Power Loss of 2-Stage	Prob of Early Stop
		Single Stage	2-Stage		
Delayed Response	18 mths	0.8275	0.7673	0.0602	0.1916
	21 mths	0.8275	0.7882	0.0393	0.2211
	24 mths	0.8275	0.8003	0.0242	0.3589
	27 mths	0.8275	0.8125	0.0150	0.4947
Changing Hazards	18 mths	0.7875	0.7423	0.0450	0.1160
	21 mths	0.7875	0.7543	0.0330	0.3011
	24 mths	0.7875	0.7600	0.0210	0.4489
	27 mths	0.7875	0.7710	0.0160	0.5941

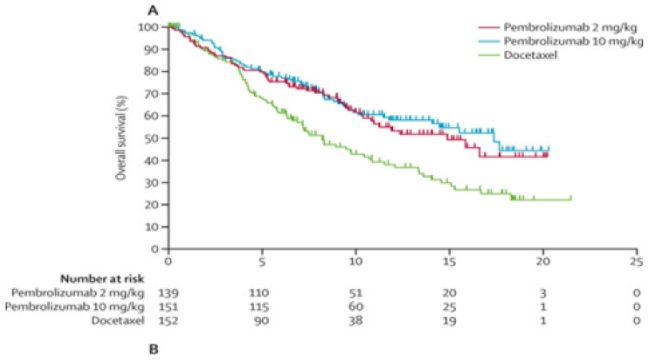
do we believe that it is advisable. By stopping early for futility there is a possibility of missing an effective therapy that might have produced a separation of the survival curves at a later time point,



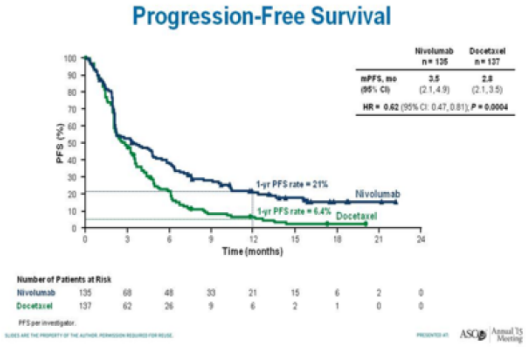
OS: Nivo in melanoma



OS: Nivo in NSCLC



OS: Pembro in NSCLC



PFS: Nivo in NSCLC

Figure 7: Kaplan-Meier Plots of Recent Immuno-oncology Trials

References

Public worshop: Oncology clinical trials in the presence of non-proportional hazards. *Duke-Margolis Center for Health Policy, Washington DC. February 2008*.

Thomas R Fleming and David P Harrington. A class of hypothesis tests for one and two sample censored survival data. *Communications in Statistics-Theory and Methods*, 10(8): 763–794, 1981.

Pranab Ghosh, Lingyun Liu, P Senchaudhuri, Ping Gao, and Cyrus Mehta. Design and monitoring of multi-arm multi-stage clinical trials. *Biometrics*, 73(4):1289–1299, 2017.

Irving K Hwang, Weichung J Shih, and John S De Cani. Group sequential designs using a family of type i error probability spending functions. *Statistics in medicine*, 9(12): 1439–1445, 1990.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Theodore G Karrison. Versatile tests for comparing survival curves based on weighted log-rank statistics. *The Stata Journal*, 16(3):678–690, 2016.

J McDermott and A Jimeno. Pembrolizumab: Pd-1 inhibition as a therapeutic strategy in cancer. *Drugs of today (Barcelona, Spain: 1998)*, 51(1):7–20, 2015.

König F Posch M Schüler A Wassmer G Ristl R, Götte H. Delayed treatment effects, treatment switches and heterogeneous patient populations: how to design and analyze rcts in oncology. *Medical University of Vienna - Merck, Poster 2019*.

Anastasios A Tsiatis. Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association*, 77(380):855–861, 1982.