SAS® GLOBAL FORUM 2020 #SASGF

Paper 5062-2020

Combination weighted log-rank tests for survival analysis with non-proportional hazards

Andrea Knezevic & Sujata Patil, Memorial Sloan Kettering Cancer Center

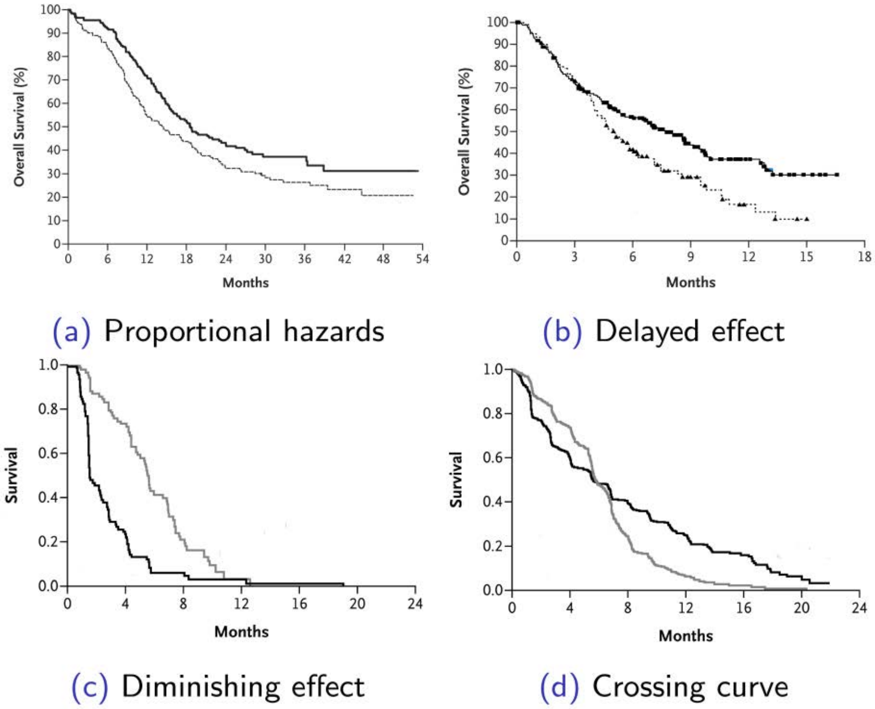
ABSTRACT

The statistical methods most commonly used to test the equality of survival curves in time-to-event analysis rely on the assumption of proportional hazards. In oncology drug development, non-proportional hazards between investigational treatments are often observed but statistical methods that properly account for these situations are rarely used in practice. The use of combinations of Fleming-Harrington weighted log-rank statistics is one relatively straightforward way to perform hypothesis testing in the presence of non-proportional hazards. In this approach, the maximum test statistic of several weighted log-rank statistics (Z_{max}) is calculated from Z-statistics obtained using the $G(\rho,\gamma)$ family. A combination test can simultaneously detect equally weighted, early, late or middle departures from the null hypothesis and can thus robustly handle several non-proportional hazard types with no a priori knowledge of the hazard functions. Although the LIFETEST procedure allows for testing with Fleming-Harrington weighted log-rank statistics, there is no built-in functionality in SAS® to test combinations of weighted tests. We discuss the development of a SAS macro that implements combination testing, including estimation of the variance-covariance matrix of the joint distribution of the Z-statistics and calculation of the p-value for Z_{max}.

INTRODUCTION

The typical approach to testing the equality of two survival curves is by using the log-rank test statistic or Cox proportional-hazards regression. Both methods work well to test the null hypothesis under the assumption of proportional hazards, or slight deviations thereof. When the two hazard functions are clearly non-proportional, the use of the log-rank test and Cox regression becomes problematic: the power of the tests to detect a difference between the curves is lost and the hazard ratio from Cox regression becomes uninterpretable. Thus, alternative analytical approaches to survival analysis are required when the assumption of proportional hazards is violated.

Figures 1 illustrates examples of the shape that two survival curves take under proportional hazards and several types of non-proportional hazards often encountered in clinical data.



(a) Proportional hazards

(b) Delayed effect

(c) Diminishing effect

(d) Crossing curve

Figure 1. Type of non-proportional hazards. (a) (Vermorken, 2007); (b) (Ferris, 2016); (c, d) (Mok, 2006).

In the delayed effect scenario, the treatment does not immediately take effect such that a lag time observed, followed by diverging hazard functions later in the follow-up period. Conversely, the treatment may provide short-term benefit and lose its effect in later follow up, resulting in converging hazard functions and a diminishing effect. Lastly, a treatment may result in higher event rate early on but may provide benefit over the entirety of the follow up period, resulting in crossing hazard functions.

The log-rank test has maximum power under proportional hazards, and in any of these non-proportional hazard scenarios the test may not detect a difference in the survival curves, especially with smaller sample sizes. For example, when hazard functions cross, events happen earlier in one group and later in the other. The log-rank scores will be positive early and negative later, so that the test statistic based on the total score may be close to zero and may not be significant, even though the two survival distributions are different. Furthermore, the estimate of the hazard ratio from Cox regression (δ) assumes that the ratio of the hazards is a constant, that is: $\delta = \frac{h_1(t)}{h_2(t)}$, for all times t, and when the ratio changes significantly over time, the value of δ becomes meaningless.

One important example of non-proportional hazards in the oncology setting is the delayed effect often seen in trials of immunotherapies. Immunotherapies quickly entered the mainstream of cancer care after their introduction in the late 1990s and they are the subject of an ever-increasing number of clinical trials (Sliwkowski & Mellman, 2013). They work by eliciting anticancer immune responses and the delayed effect is due to an indirect mechanism of action which requires time for activation of the immune system and development of an antitumour response, with a subsequent impact on clinical outcome (Hoos, 2012). Because of the strong pattern of delayed separation of survival curves in immunotherapy trials, the proportional hazard assumption is violated such that the log-rank test suffers from significant loss of power. Incorporating knowledge of the non-proportionality of hazards is essential to achieving properly powered clinical trials and appropriately interpreting trial results.

In situations where a non-proportional hazards pattern can be determined *a priori*, weighted log-rank tests can be used to test early, late or middle differences between two survival curves. However, there are many situations where investigators may not be able to predict the shape of the survival curves or whether non-proportional hazards will be observed. In these situations, combinations of weighted log-rank tests can be used to a wide range of scenarios. In this paper, we will describe the implementation of a combination test using Fleming-Harrington weighted log-rank statistics.

LOG-RANK AND WEIGHTED LOG-RANK STATISTICS

The log-rank test statistic calculates the difference in observed versus expected failures over time. Here we show the formulation of the test for the 2-sample case, which can be generalized to more than 2 samples (Kalbfleisch & Prentice, 1980). The test statistic is:

$$\chi^2 = \frac{[\sum_{t=1}^{D}(o_t - e_t)]^2}{\sum_{t=1}^{D} v_t}$$

where: $o_t = d_{1t}$, observed number of deaths in group 1 at time t,

$e_t = n_{1t}\left(\frac{d_t}{n_t}\right)$, expected number of deaths in group 1 at time t,

$v_t = d_t n_{1t}\left(\frac{n_t - n_{1t}}{n_t^2}\right)\left(\frac{n_t - d_t}{n_t - 1}\right)$, variance of expected number of deaths in group 1 at time t,

d_t, total number of deaths at time t,

n_t, total number at risk at time t, for each event time $t = 1, \ldots, D$.

A weighted log-rank test incorporates a weight function w_t that may change over time, allowing for the testing of differences between the survival curves under alternatives that differ from proportional hazards.

$$\chi^2 = \frac{[\sum_{t=1}^{D} w_t(o_t - e_t)]^2}{\sum_{t=1}^{D} {w_t}^2(v_t)}$$

Consider $\hat{S}(t-)$, the left-continuous Kaplan-Meier estimate of the survival function at time *t* for the pooled survival data. Fleming & Harrington (1982) introduced the G^{ρ} family of statistics, where $w(t) = \{\hat{S}(t-)\}^{\rho}, \rho \geq 0$. When ρ>0, early events (where $\hat{S}(t-)$ is closer to 1) are up-weighted and later events (where $\hat{S}(t-)$ is closer to 0) are down-weighted. When ρ=0, the test is equivalent to the log-rank test.

Fleming & Harrington (1991) extended this definition to the $G^{\rho,\gamma}$ family of statistics, $w(t) = \{\hat{S}(t-)\}^{\rho}\{1-\hat{S}(t-)\}^{\gamma}, \rho \geq 0, \gamma \geq 0$, which allows for the simultaneous weighting of early and late events. For example, consider ρ=0, 1 and γ=0, 1 as shown in Table 1.

ρ, γ	$w(t)$	Type of test
0, 0	1	Log-rank
1 ,0	$\{\hat{S}(t-)\}$	Test early difference
0, 1	$\{1-\hat{S}(t-)\}$	Test late difference
1, 1	$\{\hat{S}(t-)\}\{1-\hat{S}(t-)\}$	Test middle difference

Table 1. $G^{\rho,\gamma}$ family of statistics

COMBINATION TESTS

In most situations, non-proportional hazards cannot be prespecified. In these cases, a versatile test that is sensitive to both proportional hazards and a range of non-proportional hazards is desirable. One approach is to consider is combinations of weighted log-rank statistics. Combination tests aim to have good power to detect a difference in survival curves over a range of possible alternative hypotheses, which allows for testing of differences without making assumptions about the shapes of the hazard functions.

There are several examples of proposed combination tests of $G^{\rho,\gamma}$ statistics in the biometrical literature. Lee (1996) uses the combination of ($G^{0,0}$, $G^{2,0}$, $G^{0,2}$, $G^{2,2}$) to simultaneously test equally weighted, early, late and middle differences and shows that this combination has robust performance under different types of alternative hypotheses. Karrison (2016) considers the combination of ($G^{0,0}$, $G^{1,0}$, $G^{0,1}$) and provides Stata software to test any trivariate $G^{\rho,\gamma}$ combination. In the statistical literature, others have proposed more complex approaches to using weighted log-rank statistics including function-indexed statistics that simultaneously consider a large collection of values for ρ and γ (Kosorok & Lin, 1999) and tests based on weights able to adapt to changing hazard ratios (Yang & Prentice, 2010).

In 2018, the Food & Drug Administration initiated a working group with pharmaceutical companies to address issues in analysis of survival data with non-proportional hazards in the context of oncology clinical trials. The group held a public workshop to discuss and present their findings, wherein they propose the combination of ($G^{0,0}$, $G^{1,0}$, $G^{0.1}$, $G^{1,1}$) statistics, which they call the "max-combo" test (Lin, 2020). R software was developed to perform the combination test, with the option to specify any set of weights.

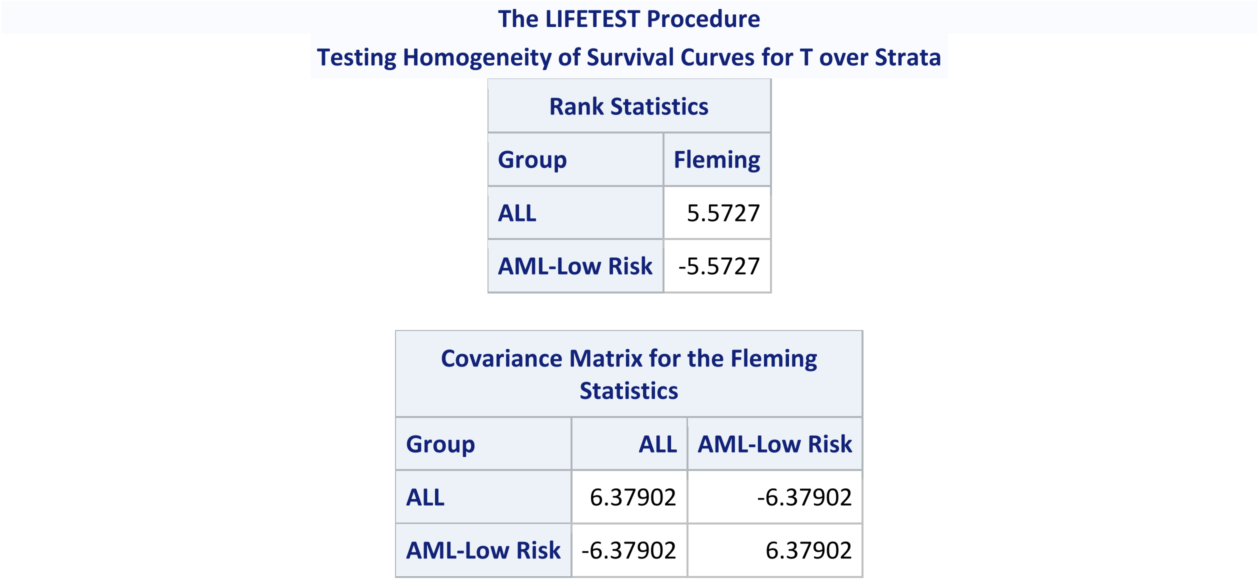
IMPLEMENTATION OF COMBINATION TESTS IN SAS

To implement the combination test, we first calculate the $G^{\rho,\gamma}$ statistics under consideration using the built-in test=FH option in the strata statement of the LIFETEST procedure. Consider two groups from the BMT dataset available in SAS to illustrate the code and output:

```
data bmt2;
 set sashelp.bmt(where=(group in ("ALL","AML-Low Risk")));
run;

proc lifetest data=bmt2;
 time T*status(0);
 strata group / test=FH(1,0);
run;
```

The test option above gives the following output with results for the $G^{1,0}$ statistic.



The LIFETEST Procedure

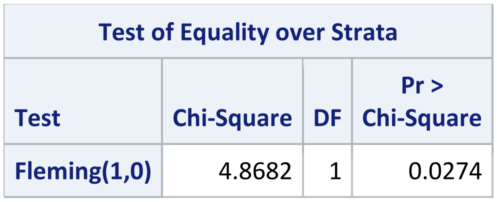
Testing Homogeneity of Survival Curves for T over Strata

Rank Statistics

Group	Fleming
ALL	5.5727
AML-Low Risk	-5.5727

Covariance Matrix for the Fleming Statistics

Group	ALL	AML-Low Risk
ALL	6.37902	-6.37902
AML-Low Risk	-6.37902	6.37902



Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Fleming(1,0)	4.8682	1	0.0274

We use the square root of the absolute value of the Chi-square estimate (Z-statistic) from the Test of Equality over Strata table to calculate Z_{max}, simply as the maximum of the Z-statistics. We will also use the variance estimates of the Z-statistics from the Covariance Matrix for the Fleming Statistics table in the p-value calculations for the combination test. Note that LIFETEST must be run multiple times to obtain results for different $G^{\rho,\gamma}$ statistics.

After calculating the necessary Z-statistics and their variance estimates, we use the same option in LIFETEST to calculate the covariance estimates between $G^{\rho,\gamma}$ statistics to complete the variance-covariance matrix. This is straightforward because, as Karrison (2016) shows, $Cov(G^{\rho1,\gamma1}, G^{\rho2,\gamma2}) = Var(G^{(\rho1+\rho2)/2, (\gamma1+\gamma2)/2})$.

Finally, to calculate the p-value for the combination test, we take 5×10^6 random samples from a multidimensional normal distribution, using mean vector 0 and the estimated variance-covariance matrix, and calculate the number of times the samples exceed Z_{max} in any dimension. This proportion is the p-value for the combination test.

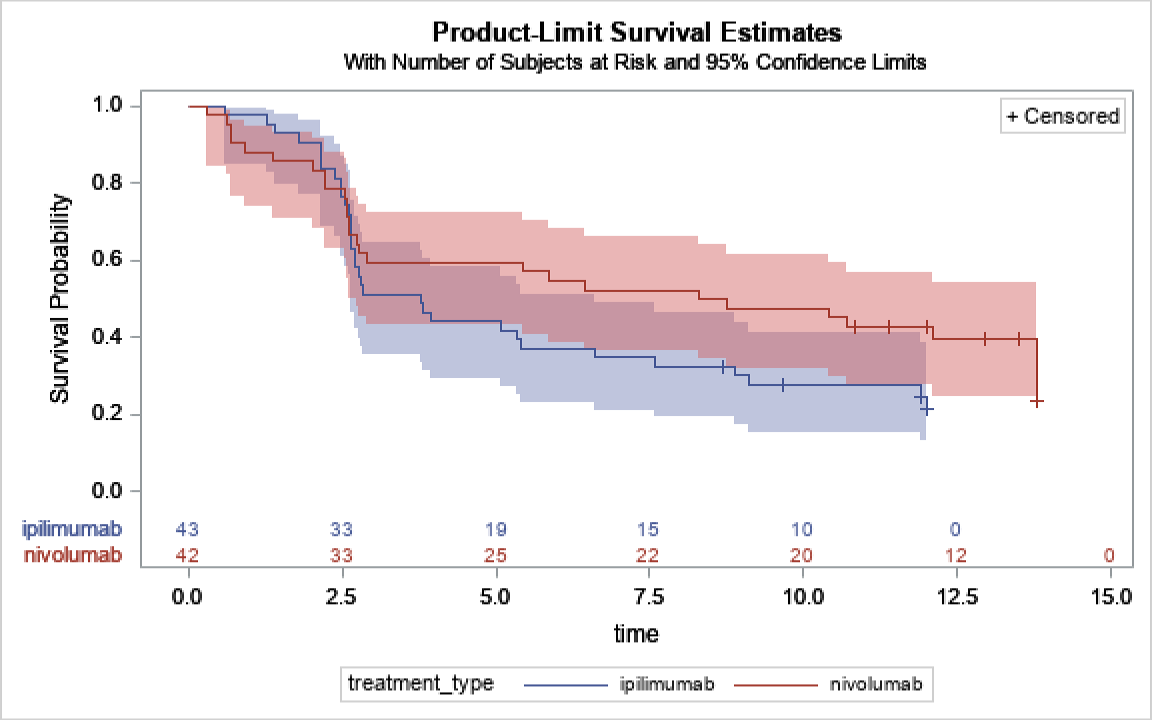
EXAMPLE COMBINATION TEST WITH NON-PROPORTIONAL HAZARDS

The following example uses digitally reconstructed data from a progression-free survival figure published in an immunotherapy trial in which nivolumab monotherapy, ipilimumab monotherapy and combination therapy were compared in 945 metastatic melanoma patients (randomized 1:1:1) (Satagopan, 2017). Here, we use the monotherapy groups only, and take a smaller, random sample of the data to illustrate the performance of the combination test with small group sample size. A total of 85 patients were selected, 43 in the ipilimumab group with 33 events and 42 in the nivolumab group with 27 events. The combination of weights proposed by Karrison (2016) are used: $G^{0,0}$, $G^{1,0}$, $G^{0,1}$.

The two survival curves show the delayed group separation that is typical of immunotherapy trials. The equally weighted log-rank and the early-weighted log rank tests do not show a significant difference between survival curves (p=0.10, 0.27 respectively). The late-weighted test is significant (p=0.02) and the combination test is also significant (p=0.04). Here we have an example where the log-rank test was not able to detect a difference in survival curves, but the combination test which simultaneously tested equally weighted, early and late differences we able to detect a difference:

```
%combo_wlr(data=larkin_85,
           group=treatment_type,
           time=time,
           event=event,
           weights=%str(0,0 1,0 0,1));
```

```
                 Combination weighted log-rank tests
```



```
                      Weighted log-rank tests

              Test           Z statistic         P
              Fleming(0,0)       1.63681    0.1017
              Fleming(1,0)       1.09293    0.2744
              Fleming(0,1)       2.29917    0.0215

                         Combination test

                         Z max        P
                       2.29917   0.0392
```

CONCLUSION

Testing combinations of weighted log-rank statistics offers an alternative to the conventional, equally weighted log-rank test that is more robust to detecting differences in survival curves in the presence of non-proportional hazards. This manuscript describes a flexible macro developed in SAS to easily calculate these tests. The macro is available for download via Github: https://github.com/dreaknezevic/combo-wlr.

The use of weighted log-rank tests that are more sensitive to alternatives is appealing in situations where non-proportional hazards are likely to be observed, such as in immunotherapy trials. Statistical obstacles remain to incorporating these tests into clinical trials, however, including calculating sample size and stopping boundaries.

It has been suggested that weighted log-rank tests be included as pre-specified sensitivity analyses in clinical trials to improve discovery of potential benefits of new treatments (Su, 2018). However, the use of combination tests has also been cautioned by authors who note that these tests can reject the null hypothesis both in favor of one group and the other on the same data and suggest that these tests risk identifying statistically significant results that are not clinically significant (Karrison, 2016; Freidlin & Korn, 2019). The role of weighted log-rank and combination tests in the future design and analysis of cancer trials remains to be seen.

REFERENCES

Vermorken JB, Remenar E, van Herpen C, et al. 2007. Cisplatin, fluorouracil, and docetaxel in unresectable head and heck cancer. *New England Journal of Medicine,* 357(17):1695-704.

Ferris RL, Blumenschein G, Fayette J, et al. 2016. Nivolumab for recurrent squamous-cell carcinoma of the head and neck. *New England Journal of Medicine,* 375(19):1856-67.

Mok TS, Wu YL, Thongprasert S, et al. 2009. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *New England Journal of Medicine*, 361(10):947-957.

Klein JP & Moeschberger ML. 2003. Survival analysis: Techniques for censored and truncated data (2nd ed.). New York, NY: Springer-Verlag.

Sliwkowski MX & Mellman I. 2013. Antibody therapeutics in cancer. *Science*, 341(6151):1192–1198.

Hoos A. 2012. Evolution of end points for cancer immunotherapy trials. *Annals of Oncology*, 23(Supplement 8):viii47–viii52.

Kalbfliesch JD & Prentice RL. 1980. The statistical analysis of failure time data. New York, NY: John Wiley & Sons.

Harrington DP & Fleming TR. 1982. A class of rank test procedures for censored survival data. *Biometrika,* 69:553-566.

Fleming TR & Harrington DP. 1991. Counting Processes and Survival Analysis. New York, NY: John Wiley & Sons.

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

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

Kosorok MR & Lin CY. 1999. The versatility of function-indexed weighted log-rank statistics. *Journal of the American Statistical Association*, 94(445):320-332.

Yang S & Prentice R. 2010. Improved log rank-type tests for survival data using adaptive weights. *Biometrics,* 66:30-38.

Duke University, US Food and Drug Administration: Public workshop: Oncology clinical trials in the presence of non-proportional hazards. https://healthpolicy.duke.edu/events/public-workshop-oncology-clinical-trials-presence-non-proportional-hazards.

Lin RS, Lin J, Roychoudhury S, et al. 2020. Alternative Analysis Methods for Time to Event Endpoints under Non-proportional Hazards: A Comparative Analysis. *Statistics in Biopharmaceutical Research*.

Satagopan JM, Iasonos A, Kanik JG. 2017. A reconstructed melanoma data set for evaluating differential treatment benefit according to biomarker subgroups. *Data in Brief*, 12:667-675.

Su Z & Zhu M. 2018. Is it time for the weighted log-rank test to play a more important role in confirmatory trials? *Contemporary Clinical Trials Communications,* 10:A1-A2.

Freidlin B & Korn EL. 2019. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? *Journal of clinical oncology*, 37(35):3455-3459.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Andrea Knezevic
Memorial Sloan Kettering Cancer Center
knezevia@mskcc.org
https://github.com/dreaknezevic