Log-rank Test: When does it Fail

- and how to fix it

Mai Zhou

Department of Statistics, University of Kentucky

- **Log-rank test**: One of the three pillars of modern Survival Analysis

(the other two are Kaplan-Meier estimator and Cox proportional hazards regression model)

- Most commonly used test to compare two or more samples nonparametrically with data that are subject to censoring.

- Quote from New England Journal of Medicine: (Jan. 5 2006)

The median duration of overall survival in the intravenous-therapy and intraperitoneal-therapy groups was 49.7 and 65.6 months, respectively (P=0.03 by the log-rank test).

Furthermore, log-rank test is the same test as the "score test" from the Cox proportional hazard model. The key words "Log-rank" and "Cox model" together appears more than 100 times in the NEJM in the last year.
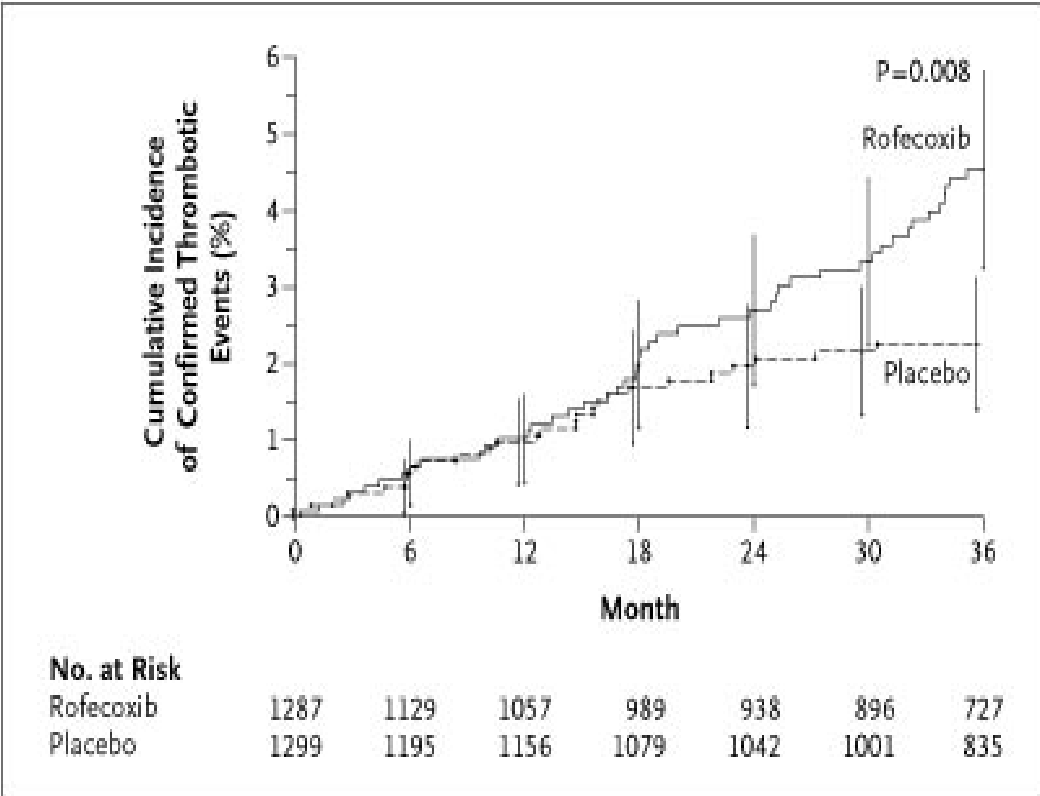
The APPROVe trial for Vioxx. See Bresalier RS, Sandler RS, Quan H, et al. (2005) NEJM

Lagakos (2006) discussed 3 issues in the statistical analysis of the trial. One of them is the proportional hazards assumption for the log-rank test and the Cox model.

There might be some evidence of non-proportionality. (P=0.07)

But no alternative test were suggested in case of cross hazard.

In general, fewer statistical procedures are available outside of proportional hazards assumption.

P=0.008

Rofecoxib

Placebo

No. at Risk							
Rofecoxib	1287	1129	1057	989	938	896	727
Placebo	1299	1195	1156	1079	1042	1001	835

- **It can fail completely.**

- Often it is used without checking appropriateness.

When does it fail? and

What are the available alternatives?

What is a log-rank test?

Consider two teams of m and n players (boys vs. girls).

1. All the players may begin a video game by putting down $1 on the table. (the ticket price).

For now suppose all players start playing at the same time.

2. When the first player fails among the $n + m$, his/her $1 on the table will be divided equally among all the players (including himself/herself), and he/she will be disqualified from further competition and leave the room with $ 1/(n+m).

3 In general, when the kth player fails, his/her \$1 on the table will be divided equally among all the players **that are active playing at the time**, including himself/herself, and then he/she will be disqualified from further competition and leave the room with his/her earned money.

The total earnings of the girl's team **is the log-rank statistic.**

(If the total net earnings of the girls team is ≈ 0 then there is no significant difference between the two teams.)

- We can show: If girls and boys are equally good at the video game, **the expected total earnings for the girls team (the \$1 she has to pay counts as negative earning) is zero**. (hint: consider each girl's individual earnings)

Regardless of censoring pattern, as long as it is independent of winning.

The competition may stop at any time, or end when all the players fail.

The game may stop when people are still actively playing. (= force those to censor; = study ends with those patients still alive.) In this case, the $1 on the table returns to those player that are censored (have not failed).

Some remarks:

● . At any time if a player wants to quit (=censoring) before been "killed in the game", it is **'fair'** to let him/her take the $1 on the table back and keep all his/her current earnings. (in fact it is a so called *martingale in time t*).

Where **'fair'** means, if he/she keeps playing, his/her expected **future** earnings is $1, so let him/her grab the $1 and quit is fair.

Proof of fairness:

think about the case with only 2 player left.

More Remark:

●. At any time if someone wants to join the game, all he/she needs to do is to put $1 on the table and start playing (just like starting a new game). In fact a player can get in or out of the game multiple times and the game is still fair, (assume he/she cannot see into future).

This is called "late entry", or "switch treatment" or more technically "time-change covariate".

(play for the girls team for a while, quit, and later re-join to play for the boys team = switch treatment).

Other interpretations of the log-rank test:

(1) sequence of 2x2 tables (Mantel-Haenszel test),

(2) weighted difference of hazard functions

(3) observed minus expected number of failures

(4) Cox model score test

(5) linear rank test

In SAS you can either use `proc phreg` (to get Cox model score test) or `proc lifetest` (to get Mantel-Haenszel test or (3)), they may be slightly different (give slightly different p-values) due to the different variance estimators used.

When no censoring, you can also use `proc npar1way savage` too.

Log-rank tests can fail if the two hazards cross.

Fail $=$ no power

Power of log-rank is best for proportional hazards type alternatives.

Some people have mixed the 'cross hazards' with 'two survival curves cross', (which can be plotted by the Kaplan-Meier estimators).

- Hazards cross and survivals cross at different places.

- Two survival functions may cross too late to show in data and plot.

- Hazards can still cross when survivals do not cross

Survivals cross $\rightarrow$ hazards cross

- Plotting of two hazard functions is not easy because the estimator of hazards are noisy; like the density, but worse in the tail.

Example: a surgical treatment has high risk in the short term but with better long term risk compare to a conservative treatment. (operation versus conservative treatment).

Possible fix:

(1) try to determine first if the hazards cross, then either use log-rank test or something else. (what is alternative?)

But people may not feel comfortable using a totally new test ....

and how to determine if there is a hazard cross (a big decision)? (post hoc decision)

Alternative (?)

For large sample size (like APPROVe trial), we may apply the log-rank for "short term" comparison only, or "long term" comparison only.

(discard half the data)

(2) Use a combined test of (log-rank test) + (a test designed specifically for cross hazards), without bother to determine if there is a crossing.

This way, you never abandon the log-rank, merely add to it something else. − Easier to accept by practitioners.

What is the catch? You lose a little power if the two hazards are actually proportional.

I will illustrate the second approach. with one example and some simulations (computation by R).

Yang and Prentice (2005) model the changing hazards ratio over time.

We simply aim to detect the difference in a test here.

How to combine two tests? (Zhou, Bathke, Kim 2006)

Basically, log-rank test is looking at a weighted difference of two hazards. Under null hypothesis, this difference is zero.

Often, other tests can also be written as a weighted difference of hazards. (different weights than the log-rank).

A test designed specifically for cross hazard alternative:

test statistic

$$\int W(t) \frac{R_1(t)R_2(t)}{R_1(t)+R_2(t)} (d\hat{H}_1(t) - d\hat{H}_2(t))$$

where $W(t)$ is a smoothed version of the function $2[I_{[t \leq T]} - 0.5]$, i.e. $= +1$ for $t \leq T$ and $= -1$ for $t > T$.

We call T the hazard cross point.

Without the $W(t)$ the above statistics is the usual log-rank test.

Use (empirical) likelihood ratio test with two constraints (both differences are zero).

Compare with chi-square 2 degrees of freedom to find p-value.

Potential problem:

Need to specify a (approx.) location of the hazard cross.

If you have a ball park idea of the location, then it is OK.

(as examples/simulation show)

First, an example:

Klein and Moeschberger (1997) have pointed out log-rank test has little or no power for cross hazard situations. They discussed several alternative tests, and an example was given. Page 211.

A clinical trial of chemotherapy against chemotherapy combined with radiotherapy in the treatment of locally unresectable gastric cancer was conducted by the Gastrointestinal Tumor Study Group (1982).

In this trial, forty-five patients were randomized to each of the two arms and followed for about eight years. The data is found in Stablein and Koutrouvelis (1985).

Plot of Kaplan-Meier and Nelson Aalen curves.
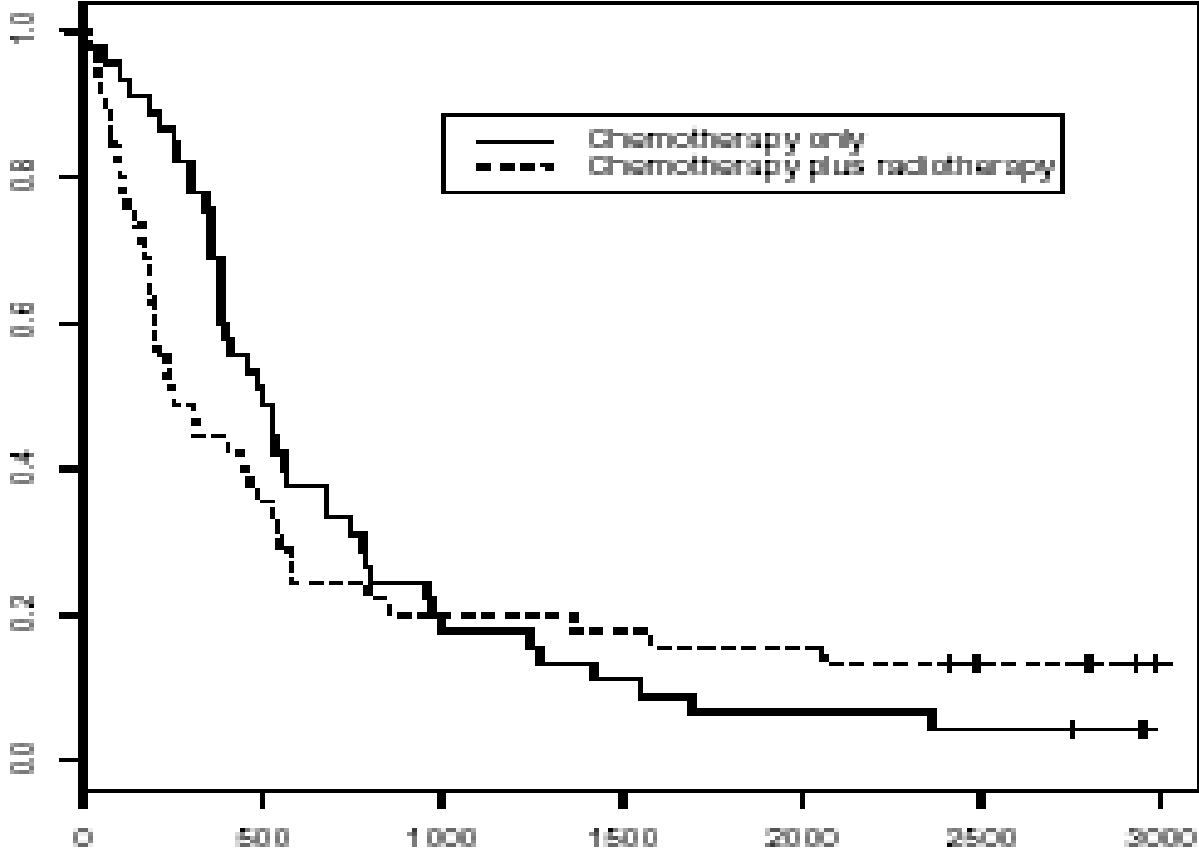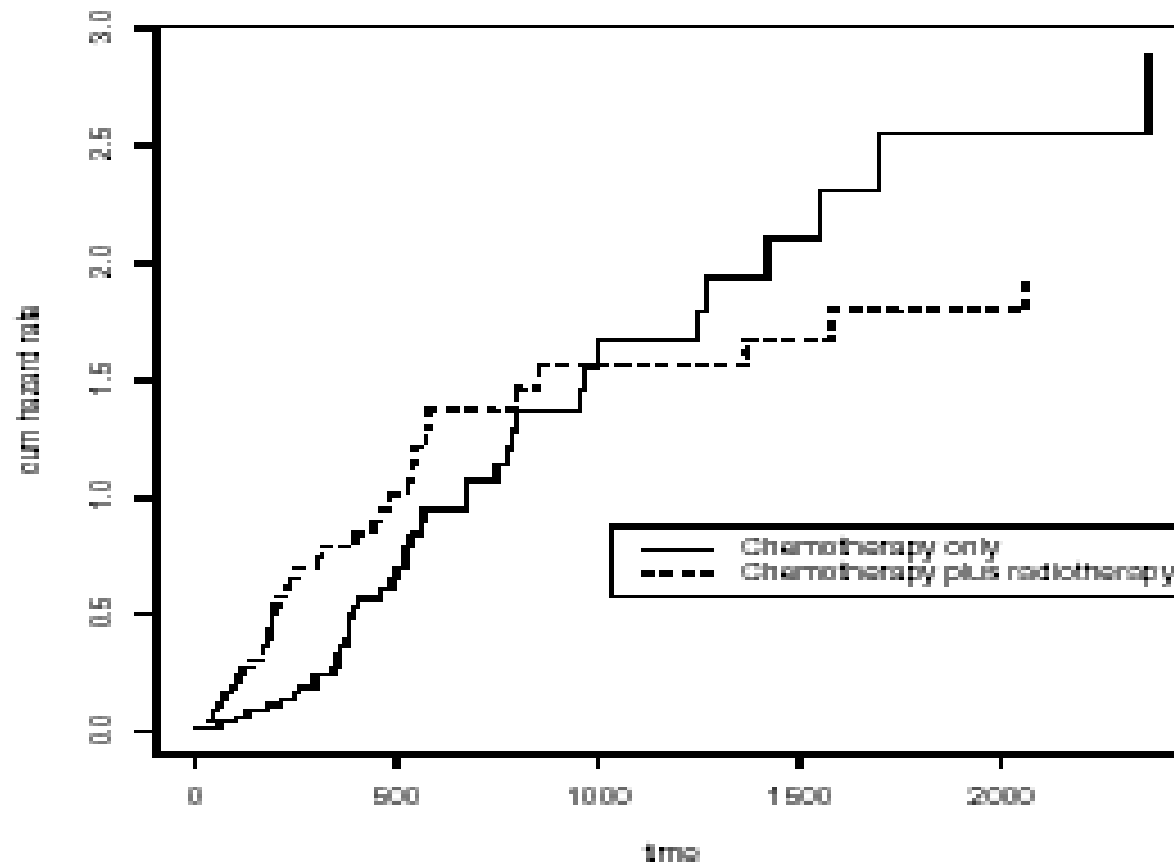
Log-rank test: P = 0.627.

Renyi type test has P = 0.053

2 Cremer von mises type tests have P = 0.06, 0.24

censored version of t-test has P = 0.74.

(above calculation done by Song Yang, a student at Wisconsin)

Next, our combined test (log-rank + test for cross hazard):

You need to pick a time of hazard crossing.

Robustness for the choice of the crossing location.

In this example chose crossing point anywhere from 150 to 1000 all give significant result: $P < 0.05$.

```
Crossing              chisq value( > 5.99 = significant)

100 ---> -2LLR = 4.4

150 ---> -2LLR = 6.1378        P=0.04647

200 ---> -2LLR = 12.559

225 ---> -2LLR = 15.234

250 ---> -2LLR = 16.2989

300 ---> -2LLR = 16.847        P=0.00022

400 ---> -2LLR = 10.97

450 ---> -2LLR = 9.429

500 ---> -2LLR = 9.455

750 ---> -2LLR = 9.955
```

850 ---> -2LLR = 8.965

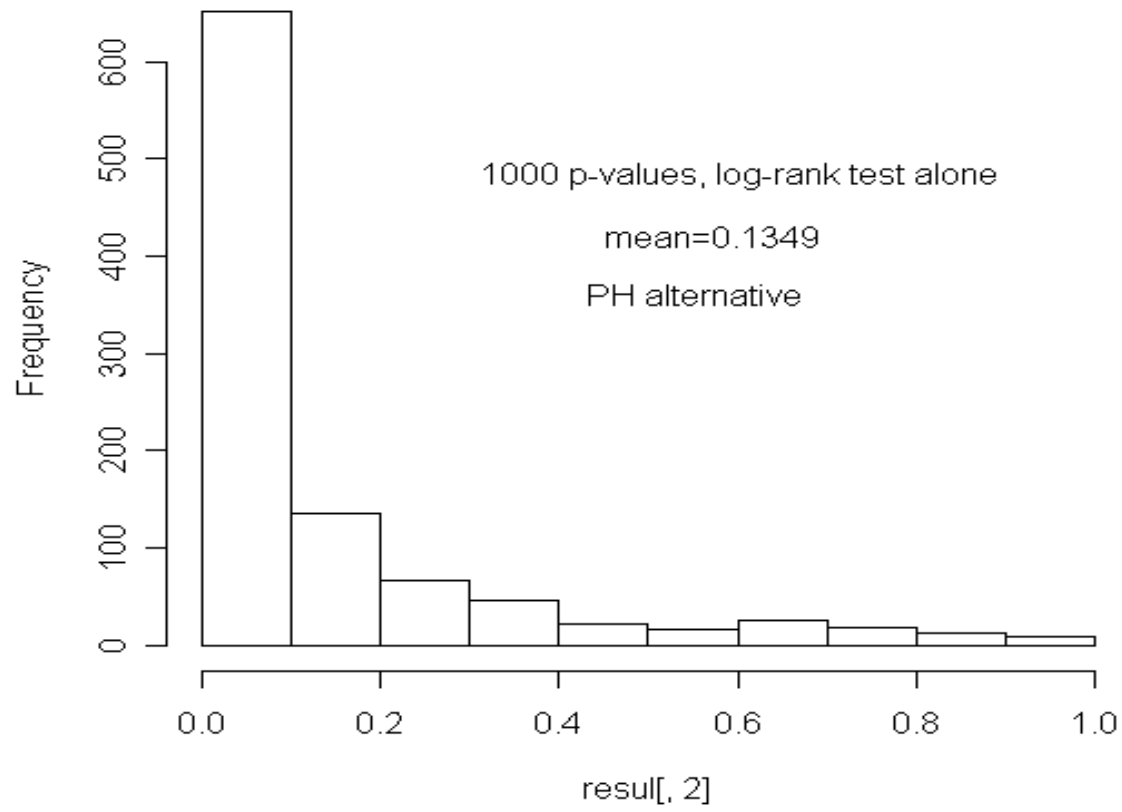1000 --> -2LLR = 6.65   P=0.03597

1100 --> -2LLR = 5.44
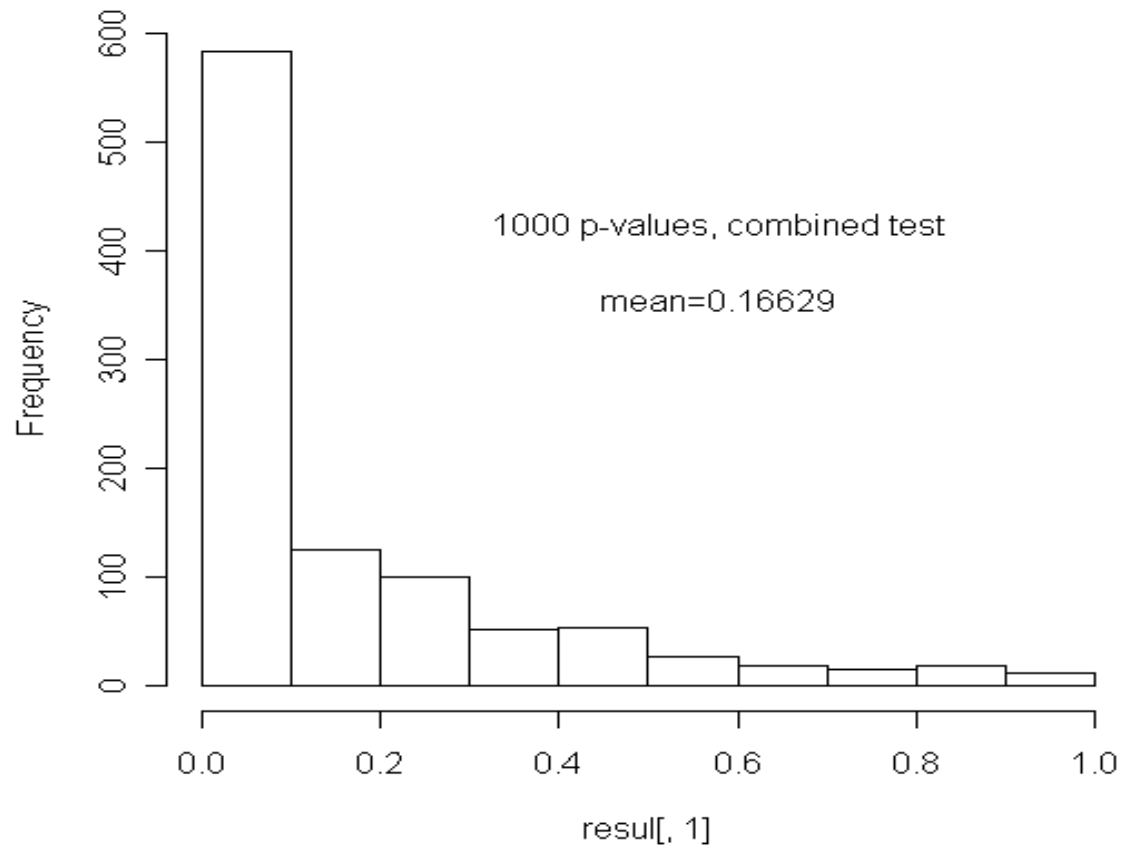
1150 --> -2LLR = 5.226

Now simulations. Case One:

Under ideal situation for log-rank, how much worse is the combined test?

Simulated 1000 tests and recorded the P-values

Histogram of resul[, 2]

1000 p-values, log-rank test alone

mean=0.1349

PH alternative

Frequency

resul[, 2]

Histogram of resul[, 1]

1000 p-values, combined test
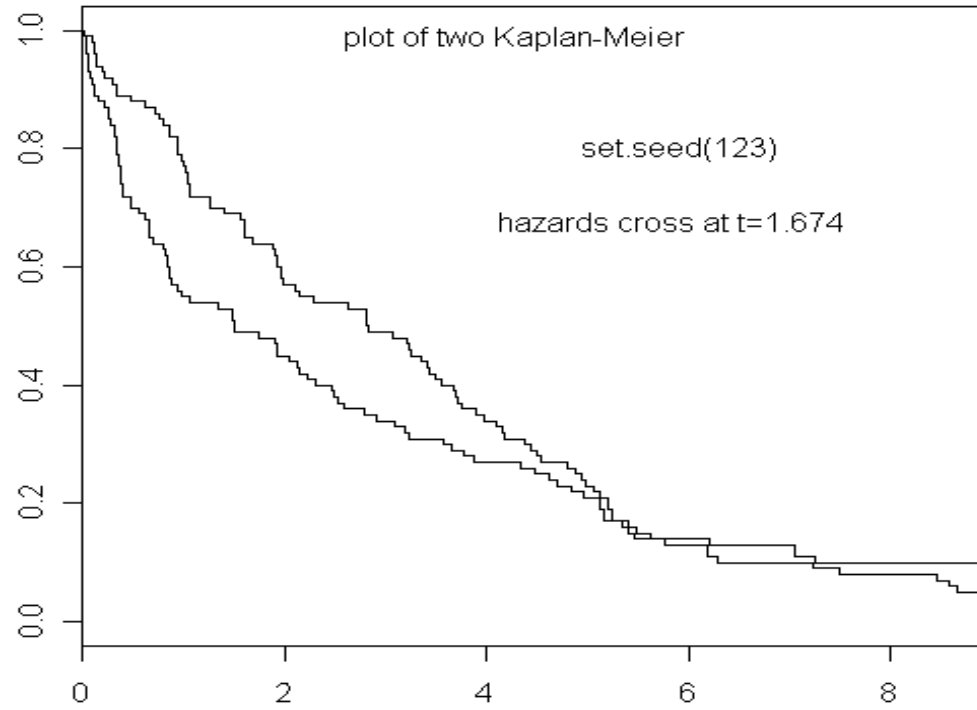
mean=0.16629

Frequency

resul[, 1]

Simulation - Case two:

Generate data from two samples: sample size of 100 each.
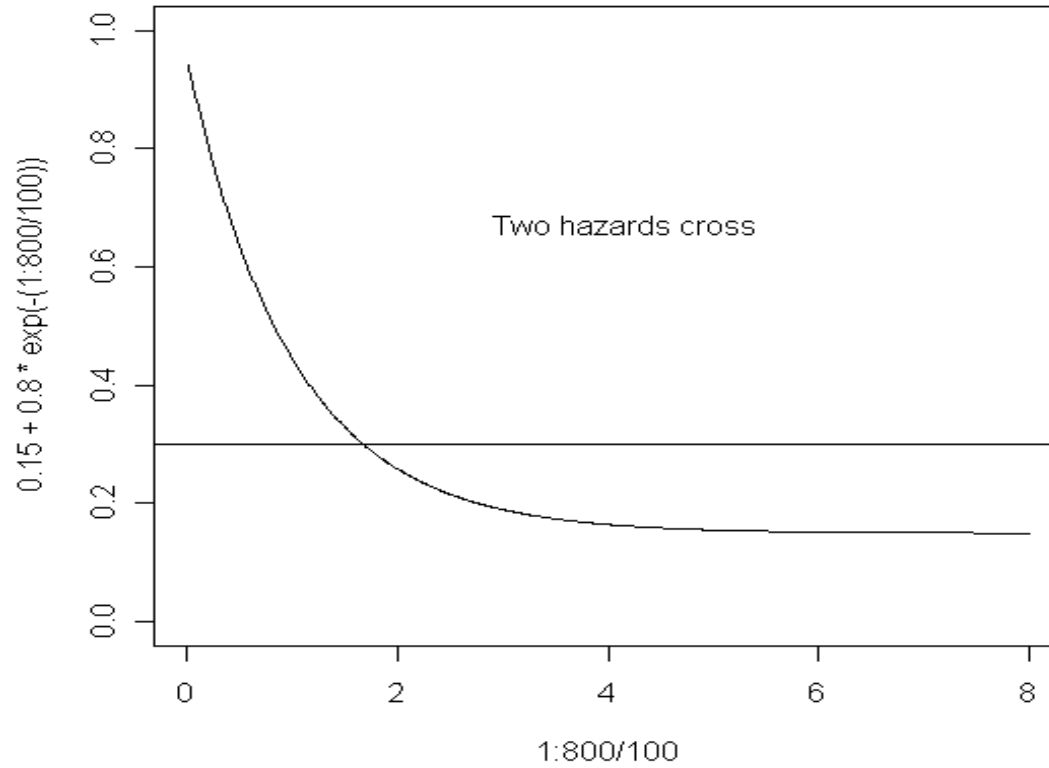
The first sample has constant hazard 0.3.

The second sample has hazard $0.15+0.8\exp(-t)$

We plot two Kaplan-Meier curves from one such data first.

plot of two Kaplan-Meier

set.seed(123)

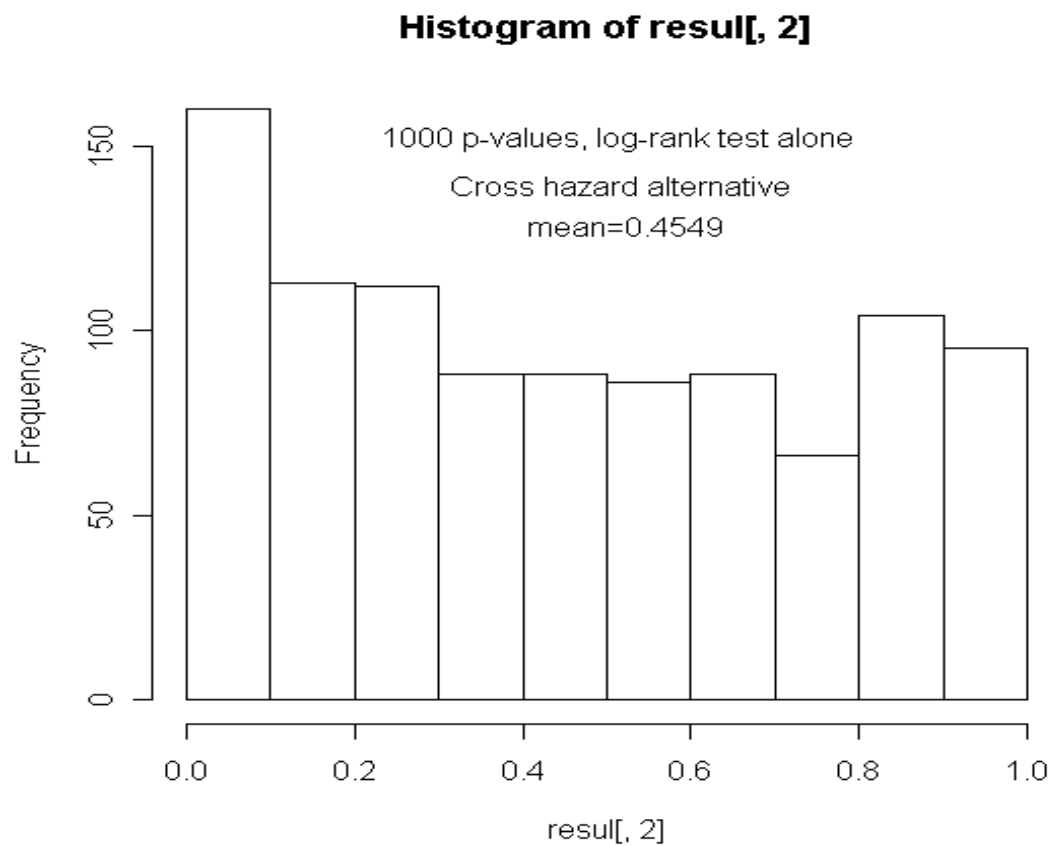hazards cross at t=1.674

and the two hazards
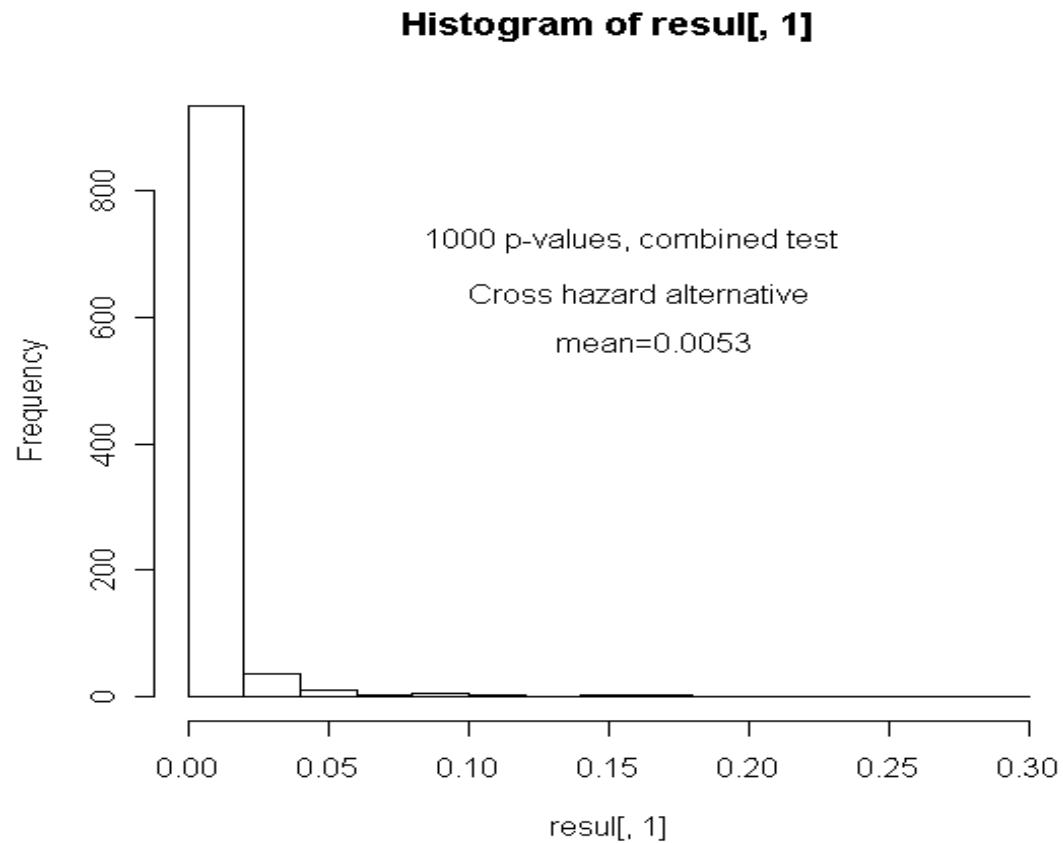
The hazard cross at $t = 1.674$.

If we censor (stop) the study at time $t < 5$, you will not see survival cross.

(diverge in the beginning, converge later)

plots of P-values: the log-rank test alone

Histogram of resul[, 2]

1000 p-values, log-rank test alone

Cross hazard alternative

mean=0.4549

Suppose we chose the crossing point T = 1.75

Histogram of resul[, 1]

1000 p-values, combined test

Cross hazard alternative

mean=0.0053

Suppose we chose the crossing point T = 1

Histogram of resul[, 1]

1000 p-values, combined test

Cross hazard alternative

mean=0.007

T = 1
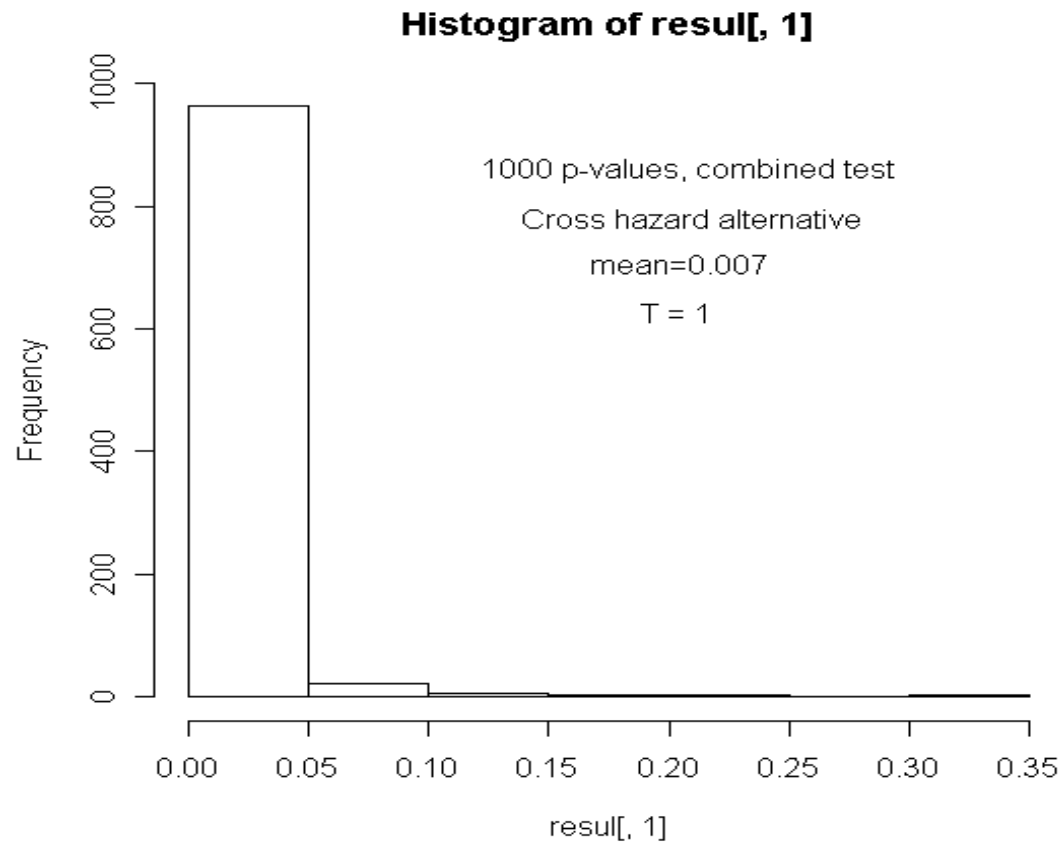
Suppose we chose the crossing point T = 2.5

Histogram of resul[, 1]

1000 p-values, combined test

Cross hazard alternative

mean=0.00825

T = 2.5

Histogram of resul[, 1]

1000 p-values, combined test

Cross hazard alternative

mean=0.0159

T = 3.5

Histogram of resul[, 1]

1000 p-values, combined test

Cross hazard alternative

mean=0.01156

T = 0.75

Histogram of resul[, 1]

1000 p-values, combined test

Cross hazard alternative

mean=0.0242

T = 0.5

Frequency
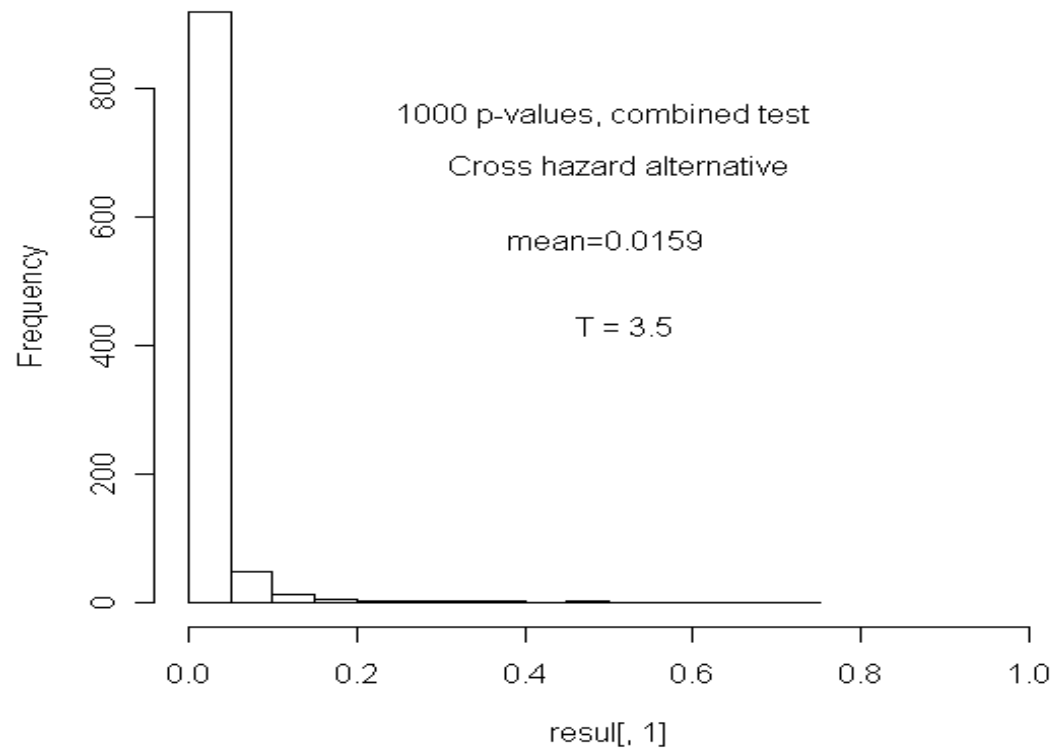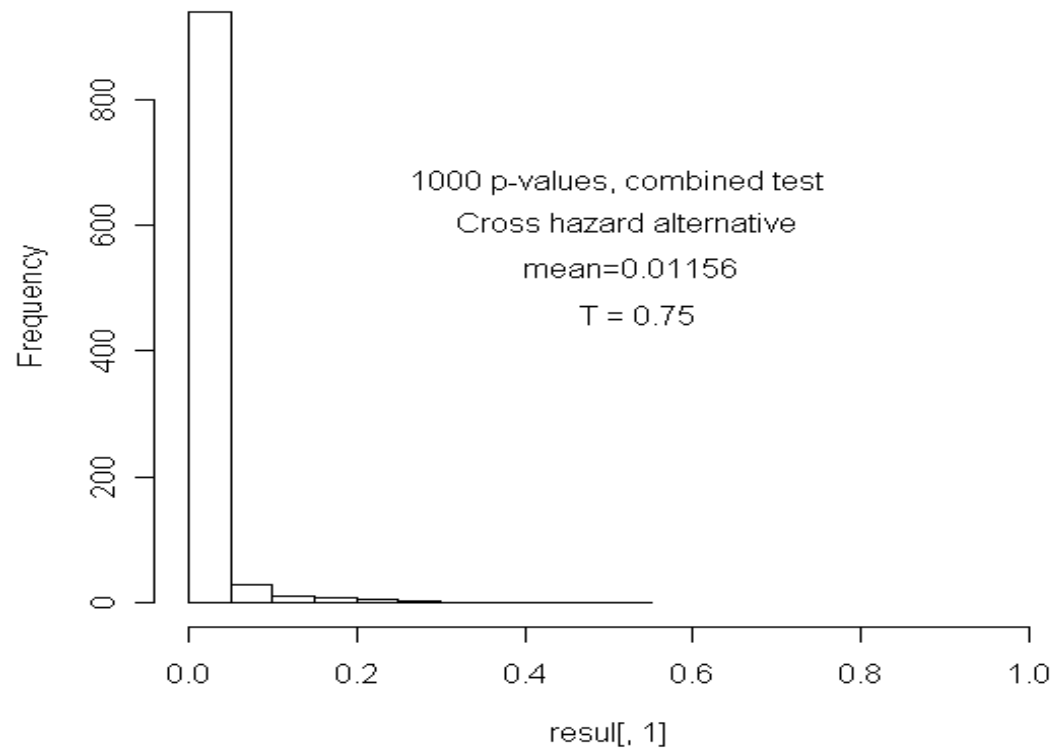
resul[, 1]

For $0.5 < T < 5$, all the combined tests are very good in power!

Conclusion

When the alternative is truly proportional hazards, the combined test lose a little power.

(Add a df that has no contribution. Look up chi-sq df=2 table instead of chi-sq df=1 table.)

When the alternative is cross hazards, the combined test is much more powerful than log-rank test.

Need to chose a (suspected) crossing point. Robust wrt the choice.

Some guidelines to chose the crossing point T:

If no info, we recommend chose the cross point at about the place where half of the expected number of failures are observed (not the median). (worst case for log-rank if the hazards cross here)

We could even chose the crossing point depend on the data, as long as it is predictable.

For situations with two or more crossing points for hazards, it may be treated similarly but not considered here.

References

Lan, G. and Lachin, J. (1995). Martingales without tears. *Lifetime Data Analysis*
   4, 361-375.

Bresalier RS, Sandler RS, Quan H, et al. (2005) Cardiovascular events associated
   with rofecoxib in a colorectal adenoma chemoprevention trial. N Engl J Med
   2005;352:1092-1102.

Lagakos, S. W. (2006) Time-to-Event Analyses for Long-Term Treatments: The
   APPROVe Trial. N Engl J Med 355, 113-117.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis − Techniques for
   Censored and Truncated Data*, Springer, New York.

Owen, A. (2001). *Empirical likelihood*. Chapman & Hall, London.

Zhou, M., Bathke, A. and Kim, M. (2006) Combined Multiple Testing by Censored
   Empirical Likelihood. Univ. Kentucky, Department of Statistics Tech Report.
   Submitted/under revision.

Song, Y. http://www.cs.wisc.edu/∼songyang/

Yang, S. and Prentice, R. (2005). Semiparametric analysis of short-term and
   long-term hazard ratios with two-sample survival data *Biometrika* **92**, 1-17.

The test function is written in R — a free statistical language similar to Splus; and the code is available at

http://www.ms.uky.edu/ mai/splus.html/WLogRk.r

> library(emplik)

> WLogRk(x1=times1, d1=status1, x2=times2, d2=status2, T=200)

sample one: constant hazard of 0.3
sample two: cumulative hazard = 0.15t + 0.8 [1- exp(-t)]
            hazard = 0.15 + 0.8 exp(-t)

for short term, sample two is worse, has 4 times hazard of sample one.
for long term, sample two has 50% lower hazard compare to sample one.

Range of r.v. in (0, 8)

Generating the random variables:

Need to find the inverse function.

```
> hazfun <- function(t, a=0.15, b=0.8) {
a*t + b*(1-exp(-t))
}
### this is cumulative hazard function.

> invhazfun <- function(y, fun=hazfun){
myhazfun1 <- function(t){ hazfun(t) - y }
temp <- uniroot(f = myhazfun1 , lower=-0.1, upper=100)
return(temp$root)
```

```
}

> rvgenerate <- function(n=1){
temp <- rexp(n)
for(i in 1:n) temp[i] <- invhazfun(temp[i])
return(temp)
}


> simuLogRank <- function(n=100) {
obs1 <- rexp(n)/0.3
obs2 <- rvgenerate(n)
x <- c( rep(1,n) , rep(2,n) )
tempout <- survdiff(Surv(c(obs1, obs2), rep(1,2*n))~ x )
return(tempout$chisq)
}


simuWLogRk <- function(n=100) {
obs1 <- rexp(n)/0.3
obs2 <- rvgenerate(n) ####  or rexp(n)/0.15
```

```
temp11 <- Wdataclean3(z=obs1, d=rep(1,100))
temp12 <- DnR(x=temp11$value, d=temp11$dd, w=temp11$weight)
TIME <- temp12$times
RISK <- temp12$n.risk
fR1 <- approxfun(x=TIME, y=RISK, method="constant", yright=0, rule=2, f=1)

temp21 <- Wdataclean3(z=obs2, d=rep(1,100) )
temp22 <- DnR(x=temp21$value, d=temp21$dd, w=temp21$weight)
TIME <- temp22$times
RISK <- temp22$n.risk
fR2 <- approxfun(x=TIME, y=RISK, method="constant", yright=0, rule=2, f=1)

flogrank <- function(t){fR1(t)*fR2(t)/(fR1(t)+fR2(t))}
myfun6 <- function(x) { temp <- 8*( 0.5 - x )
return( pmax( -1, pmin(temp, 1)) ) }
fWlogrank <- function(t) { myfun6(t/3.5)*flogrank(t) }
##### because the hazard cross at arount 3.5/2=1.75 ####
fBOTH <- function(t) { cbind( flogrank(t), fWlogrank(t) ) }

out1 <- emplikHs.test2(x1=obs1, d1=rep(1,100), x2=obs2, d2=rep(1,100),
            theta=c(0,0), fun1=fBOTH, fun2=fBOTH)
```

```
x <- c( rep(1,n) , rep(2,n) )
tempout <- survdiff(Surv(c(obs1, obs2), rep(1,2*n))~ x )
return( c(out1$"-2LLR", tempout$chisq))
}



The stand alone function:

WLogRk <- function(x1, d1, x2, d2, T, simpleTest = FALSE) {
###################################################################
### T is the crossing point of the two hazards.
### If simpleTest is TRUE, then it also returns the regular log-rank test
### P-value. Should be similar to SAS proc lifetest, R survdiff() .
###################################################################

temp11 <- Wdataclean3(z=x1, d=d1)
temp12 <- DnR(x=temp11$value, d=temp11$dd, w=temp11$weight)
TIME <- temp12$times
RISK <- temp12$n.risk
fR1 <- approxfun(x=TIME, y=RISK, method="constant", yright=0, rule=2, f=1)

temp21 <- Wdataclean3(z=x2, d=d2 )
```

```
temp22 <- DnR(x=temp21$value, d=temp21$dd, w=temp21$weight)
TIME <- temp22$times
RISK <- temp22$n.risk
fR2 <- approxfun(x=TIME, y=RISK, method="constant", yright=0, rule=2, f=1)

flogrank <- function(t){fR1(t)*fR2(t)/(fR1(t)+fR2(t))}
myfun6 <- function(x) {temp <- 8*( 0.5 - x )
                        return( pmax( -1, pmin(temp, 1)) ) }

fWlogrank <- function(t) { myfun6(t/(2*T))*flogrank(t) }
###############################################
fBOTH <- function(t) { cbind( flogrank(t), fWlogrank(t) ) }

out1 <- emplikHs.test2(x1=x1, d1=d1, x2=x2, d2=d2,
                        theta=c(0,0), fun1=fBOTH, fun2=fBOTH)

pvalue <- NA
if(simpleTest) {
out2 <- emplikHs.test2(x1=x1, d1=d1, x2=x2, d2=d2,
                        theta=0, fun1=flogrank, fun2=flogrank)
pvalue <- 1-pchisq(out2$"-2LLR", df=1)  }
```

```
list(Pval = 1-pchisq(out1$"-2LLR", df=2), Pval(log-rank) = pvalue )
}
```

Quoting Lagakos (2006) NEJM:

The second issue raised by the analysis of the cardiovascular data is that of the assumption of proportional hazards. The log-rank and Cox tests are motivated by this assumption that is, that the relative risk remains constant over time. Given this assumption, the relative risk provides a simple way of describing the magnitude of the effect of treatment on the end point, and one can infer that the corresponding cumulative incidence curves diverge throughout the entire time range covered. These tests can be well powered to detect some differences between treatment groups that do not satisfy the assumption of proportional hazards, but they can have poor power to detect other differences, including cumulative incidence curves that are initially equal but later diverge and others that initially diverge but later approach one another. When either test yields a nonsignificant difference between the treatment groups, one concern is whether the treatments could differ in a way that is not captured by the test. Thus, the proportional-hazards assumption is tested to determine whether a nonsignificant difference between groups might have been due to a treatment effect that does not satisfy that assumption.

The most common analytic way of testing the proportional-hazards assumption is by fitting a Cox model with one term representing the treatment group and another term representing an interaction between the treatment group and either time or the logarithm of time. These models correspond to a relative risk that changes exponentially (relative risk(t)=et) or as a power of time (relative risk(t)=t). Which of these two interaction tests is more powerful will depend on the nature of the difference between the treatment groups. When applying them, it is important to keep in mind that rejection of the proportional-hazards assumption does not mean that the true relative risk follows the form assumed in an expanded Cox model, nor does the failure to reject the assumption necessarily mean that the assumption holds.

The APPROVe investigators planned to use an interaction test with the logarithm of time as the primary basis for testing the proportional-hazards assumption. This test resulted in a P value of 0.07, which did not quite meet the criterion of 0.05 specified for rejecting the assumption. However, the original report of the APPROVe trial1 mistakenly gave the P value as 0.01, which was actually the result of an interaction test involving untransformed time. (This error is corrected in this issue of the Journal.) The investigators noted that the estimated cumulative incidence curves for adjudicated serious thrombotic events in the rofecoxib and placebo groups were similar for approximately the first 18 months of treatment and thereafter diverged. I interpreted this statement as no more than a simple way of describing the visual difference between the Kaplan-Meier curves for the rofecoxib and placebo groups and not as a claim that the cumulative incidence rates were equivalent in the two groups for the first 18 months, since this neither was demonstrated nor follows from the use of either of the interaction models used to test the proportional-hazards assumption.

The estimated relative risk calculated with the use of the Cox model represents a time-averaged hazard ratio and thus may not adequately describe the difference between the treatment and placebo groups when the proportional-hazards assumption does not hold. It may then be of interest to assess how the cumulative incidence curves might plausibly differ over time. Doing so by means of post hoc analyses based on visual inspection of the shapes of the Kaplan-Meier curves for the treatment groups can be misleading and should be avoided. A better approach is to create a confidence band for the difference between the cumulative incidence curves in the treatment and placebo groups that is, for the excess risk in the treatment group. Confidence bands can be constructed in several ways and for settings in which some observations are informatively censored. The bands are commonly centered around the estimated difference between the treatment groups, so that for a 95 percent band, the 5 percent error is evenly split above and below the band.

Quoting FDA web:

Merck's decision to withdraw Vioxx from the market is based on new data from a trial called the APPROVe [Adenomatous Polyp Prevention on VIOXX] trial. In the APPROVe trial, Vioxx was compared to placebo (sugar-pill). The purpose of the trial was to see if Vioxx 25 mg was effective in preventing the recurrence of colon polyps. This trial was stopped early because there was an increased risk for serious cardiovascular events, such as heart attacks and strokes, first observed after 18 months of continuous treatment with Vioxx compared with placebo.