Analytical Methods Under Non-Proportional Hazards: A Dilemma of Choice

Amarjot Kaur, Qing Li, Jing Li

Merck Research Labs

Indiana University

Regulatory-Industry Workshop (12-14 Sept 2018)

Outline

- Background
- Available (Selected) Methods
  - Testing and/or Estimation
- Simulation Studies
- Illustrative Example
- Summary

Background

For time-to-event data

- Cox proportional hazards (PH) model and Log-rank test are the commonly used methods.

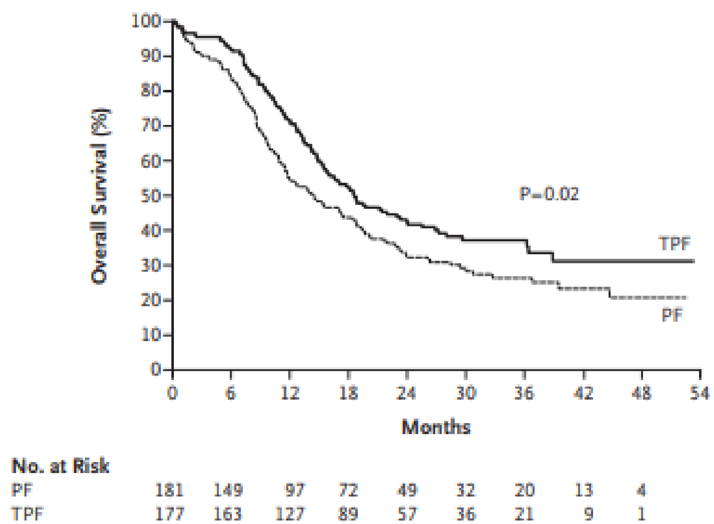
  (PH hazard ratio between two arms is constant over time)

- Results typically reported as
  - Kaplan-Meier (KM) curves, including estimated median survival time
  - Log-rank test: p-Values (testing)
  - Cox PH model: hazard ratio & p-Values (estimation & testing)

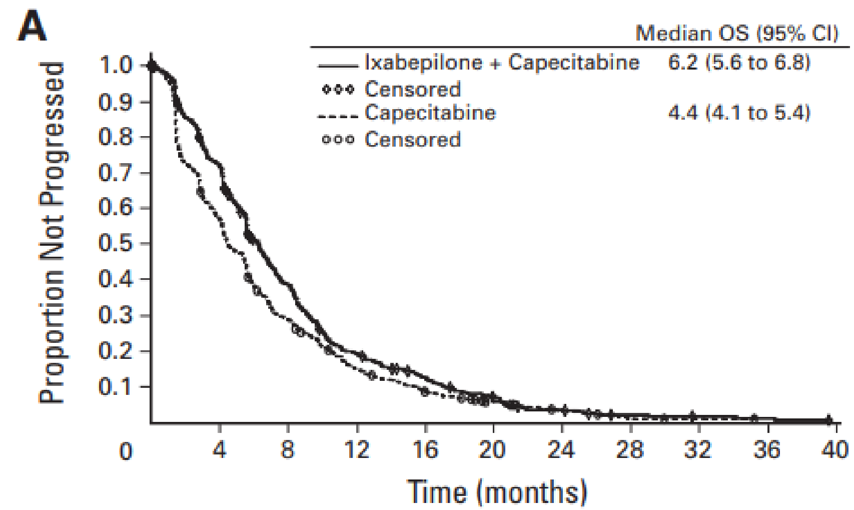
- When two hazard rates are non-proportional, the power is lost for both log-rank & Cox PH test
  - Log-rank no longer the most powerful test
  - the score test based on Cox model is no longer the best partial-likelihood statistics

Examples - KM curves for overall survival

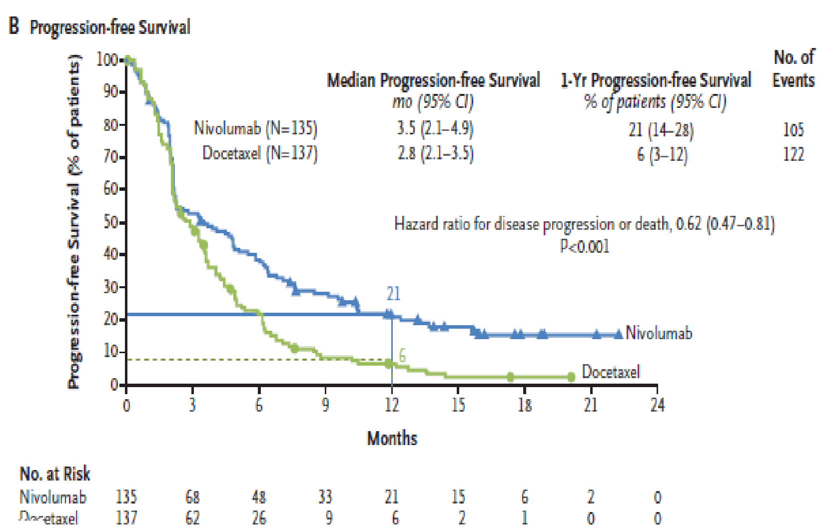
Proportional Hazards



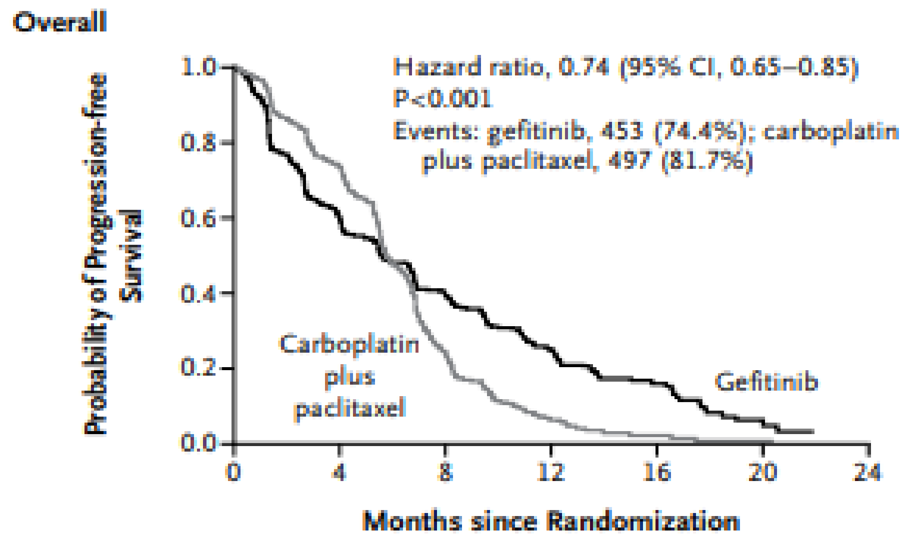
Early/Diminishing Effect



Late/Delayed Effect



Crossing Hazards



Background – Non-proportional Hazards

Type of non-proportionality

- Quantitative Interaction (Non-Crossover Interaction)
  The hazards ratio varies over time in magnitude but not in direction.
  (Cox PH model has moderate performance with mild quantitative interaction)

- Qualitative Interaction (Crossover Interaction)
  The hazards ratio varies over time with change in direction.
  (Cox PH model has substantially low performance under qualitative interaction; interpretation of test results not meaningful)

Sources of non-proportionality

- Treatment-by-time interaction
- Subgroups
- Unobservable or un-measureable random effect (frailty)

What To Do When NPH is known?

- Once the evidence of non-proportional hazards is known then the next step would be to incorporate this information in the analyses.

- NPH impacts
  - Trial design: Sample size /power analysis
  - Data analysis: Testing and estimate

- *But what method to use amongst many available?*

  Understanding the extent and source of NPH would be helpful.

Some Commonly Used Methods

- Parametric Model (Weibull, AFT, etc.)
- Piecewise Exponential Model

- Weighted Log-Rank Test
  - Log-rank with adaptive weights
- Max-Combo Test

Rank based

- Weighted Kaplan-Meier Test
- Restricted Mean Survival Time (RMST)

K-M based

- Approaches using Cox PH
  - Treatment-by-covariate interaction by including time varying covariate
  - Treatment-by-stratum interaction by combining stratum-specific estimates
  - Cox PH model with change point (HRs for two or more timeperiods)

- Other Methods
  - Renyi Type Tests
  - Gamma Frailty Model
  - More…

Weighted Log-Rank Test

Test statistic $W_{WLRT} = U/\sqrt{V}$

$$U = \int_0^\infty K(s)\frac{\bar{Y}_2(t)}{\bar{Y}(s)}d\bar{N}_1(s) - \int_0^\infty K(s)\frac{\bar{Y}_1(t)}{\bar{Y}(s)}d\bar{N}_2(s)$$

$$V = \int_0^\infty K^2(s)\frac{\bar{Y}_1(s)\bar{Y}_2(s)}{\bar{Y}^2(s)}d\bar{N}(s)$$

$*\ \bar{N}_j(s)$: # of failures at time s from group j $(j=1,2)$

$*\ \bar{Y}_j(s)$: # of subjects at risk at time s from group j $(j=1,2)$ and $\bar{Y}(s) = \bar{Y}_1(s) + \bar{Y}_2(s)$

$*\ K(s)$: for $G^{\rho,\gamma}$ statistics

$$K(s) = [\hat{S}(s-)]^{\rho}[1-\hat{S}(s-)]^{\gamma}$$

$\hat{S}$ is the Kaplan - Meier estimators for the pooled sample

Pros

- Easy to implement & offers flexibilities on choice of weight for different scenarios
- With correct choice of weight, the efficiency of this test is much better than LRT and Cox model under NPH

Cons

- Correct choice of weights is a challenge

The efficiency of this test could be very low with a improper weight

Weighted Kaplan-Meier Test

- Pepe and Fleming (1989) proposed a test for a general class of alternative:
- Test Statistic:

$$H_1 = S_1(t) \geq S_0(t) \text{ for all } t.$$

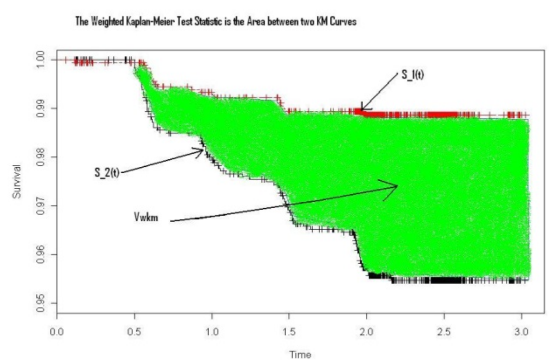
$$V_{WKM} = \int_0^{\infty} K(t) \{\hat{S}_1(t) - \hat{S}_2(t)\} \, dt$$

$$\text{where } K(t) = \frac{\hat{C}_1^{-}(t)\,\hat{C}_2^{-}(t)}{n_1/(n_1+n_2)\,\hat{C}_1^{-}(t) + n_2/(n_1+n_2)\,\hat{C}_2^{-}(t)}$$

$* \hat{S}_1(t)$ and $\hat{S}_2(t)$ are K-M estimators for the survival functions

$* \hat{C}_1(t)$ and $\hat{C}_2(t)$ are K-M estimators for censoring distribution functions

V_{WKM} is the weighted difference of area under curve (AUC) of two K-M curves; Special case of $K(t) = 1$



Pros

Concept is easy to understand

Choice of weight could be objective (e.g., only depends on censoring)

Cons

When weight is determined by censoring, the performance of the test becomes sensitive to the censoring

Weight Functions – Treatment Effect Testing

► (Weighted) log-rank tests

- Weight function

  $FH(\rho,\gamma) = \widehat{S(t)}^{\rho} \cdot \left(1 - \widehat{S(t)}\right)^{\gamma}$

- FH(0,0): log-rank test
- FH(0,1): late effect
- FH(1,0): early effect
- FH(1,1): middle effect

► Weighted Kaplan-Meier test

- Weight function

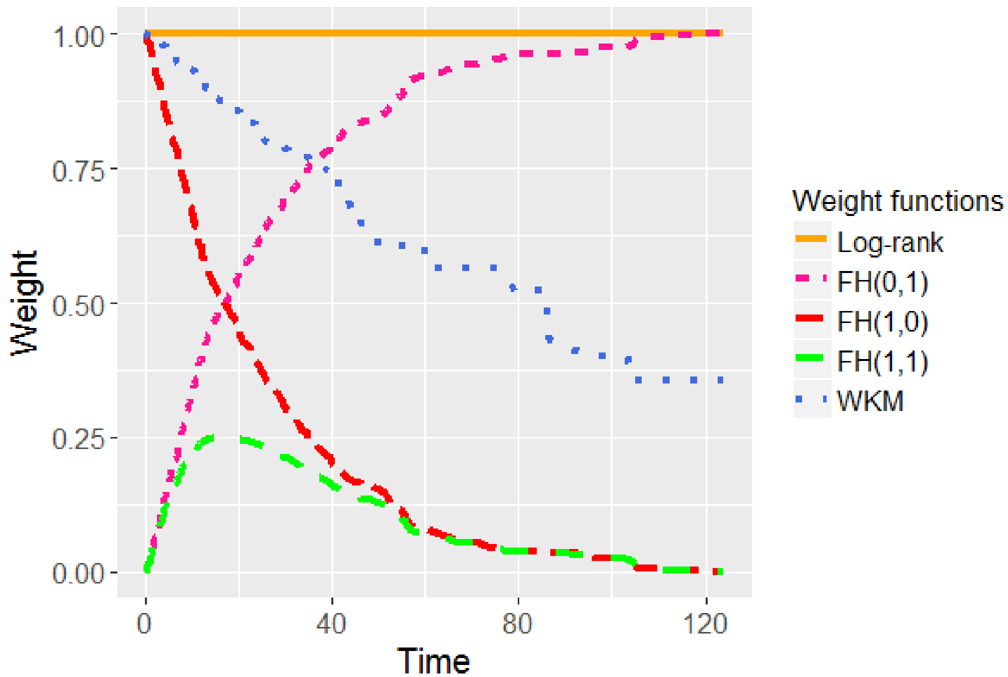
  $\widehat{w}_c(t) = \frac{\hat{C}_1^-(t)\hat{C}_2^-(t)}{\hat{p}_1\hat{C}_1^-(t)+\hat{p}_2\hat{C}_2^-(t)}$,

  where $\hat{C}_i^-(t)$ is prob of not being censored before time t

  (i.e., censoring survival function)

  weights monotonically decreasing with time

Weight Functions



PH scenario with HR=0.75

Proprietary

Restricted Mean Survival Time (RMST)

- $T_R(t) = \int_0^t S(u)du = t \times \frac{\int_0^t S(u)du}{t} = t \times \bar{S}(t)$
  - $\bar{S}(t)$: mean survival function from 0 to t
  - T_R: mean survival time from 0 to t or RMST
- Pros
  - RMST is a good point estimate under NPH comparing to HR from Cox PH model
  - RMST can easily be estimated from K-M method
- Cons
  - Requires a proper landmark time and value of point estimate can be greatly influenced by later time variability

Max-Combo Test

(FDA-Duke-Margolis NPH Workshop 2018)

A combination of *FH(ρ,γ)* weighted log-rank tests

Details

- Let Z_1, Z_2, Z_3, Z_4 be test statistics of weighted log-rank tests with weights FH(0,0), FH(0,1), FH(1,0), and FH(1,1).
- Test statistic:

$$Z_{max} = \max(|Z_1|, |Z_2|, |Z_3|, |Z_4|)$$

- Under *H0*, $(Z_1, Z_2, Z_3, Z_4) \implies MVN_4(0, \Sigma)$
  - $\Sigma = \left(\sigma_{ij}\right)_{4\times 4}$, where

    $$\sigma_{ij} = \frac{n_1+n_2}{n_1 n_2}\int_0^\infty K_l(t)\,K_m(t)\,\frac{\overline{Y_1}(\boldsymbol{t})\overline{Y_2}(\boldsymbol{t})}{\overline{Y_1}(\boldsymbol{t})+\overline{Y_2}(\boldsymbol{t})}\left(1-\frac{\Delta\overline{N_1}(\boldsymbol{t})+\Delta\overline{N_2}(\boldsymbol{t})-1}{\overline{Y_1}(\boldsymbol{t})+\overline{Y_2}(\boldsymbol{t})-1}\right)\left[\frac{d\{\overline{N_1}(\boldsymbol{t})+\overline{N_2}(\boldsymbol{t})\}}{\overline{Y_1}(\boldsymbol{t})+\overline{Y_2}(\boldsymbol{t})}\right]$$
  - Gill, 1980; Kosorok and Lin, 1999; Karrison et al., 2016
- *P*-value: derived via integration of multi-variate Normal distribution

Pros

Well-controlled type I error rate; Robust to various profiles of NPH in terms of power

Cons

Clinical justification on weight functions; Lack of coherent estimation procedure (weighted HR may not suffice)

Simulation Studies

1. To compare available methods under quantitative & qualitative interactions

   Type I error and power; one-sided vs two-sided testing?

2. To examine Cox model with change point.

Simulation set up

- N = 500 (1:1 ratio); 10,000 replications
- Data are simulated from piecewise exponential survival model.
- Independent exponential censoring

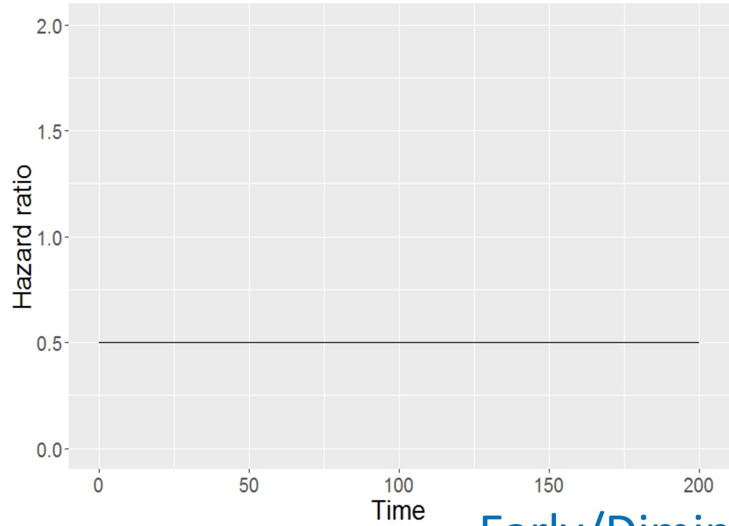
Different scenarios

- Proportional hazards (PH)
- Non-proportional hazards (NPH)
  - Early/Diminishing effect
  - Late/Delayed effect
  - Crossing hazards

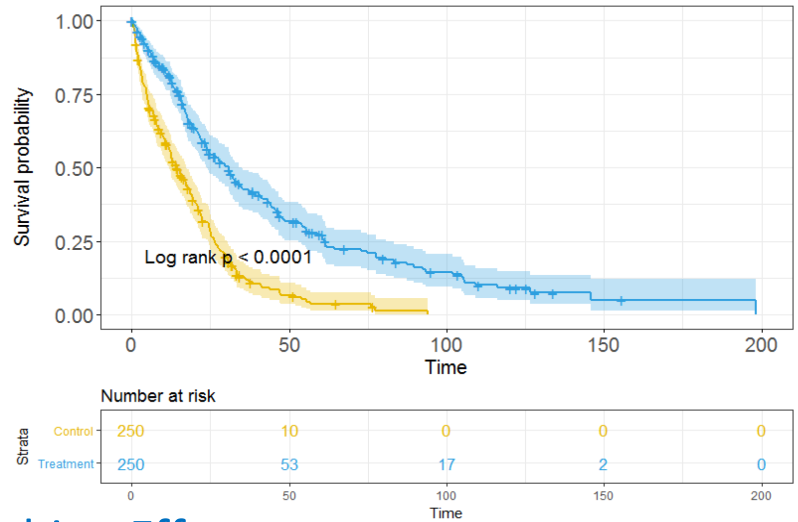
Different Scenarios (non-crossing hazards)

Proportional Hazards

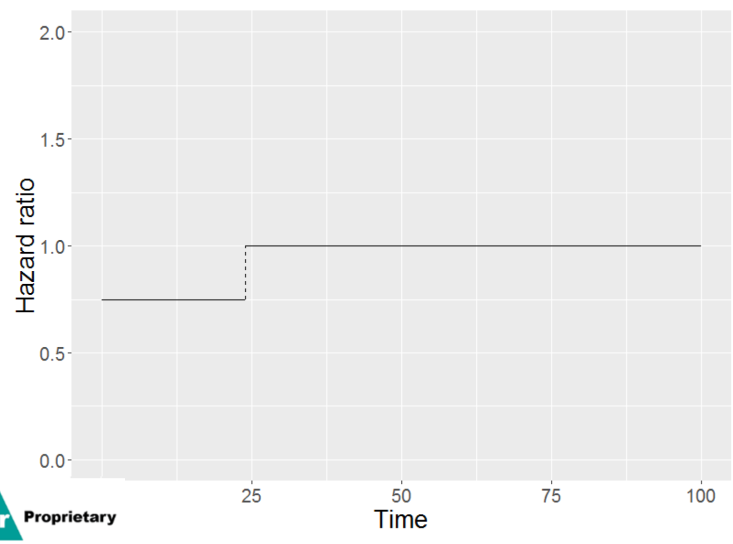
Hazard Ratio

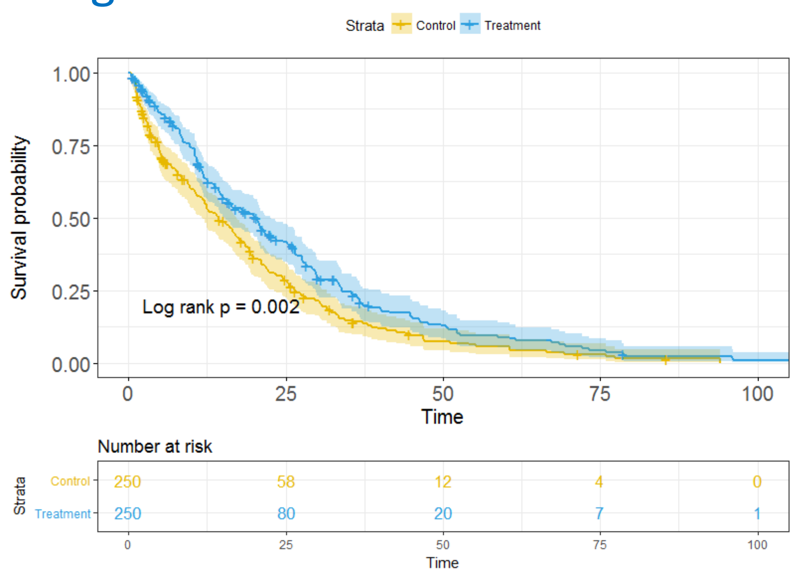


Kaplan-Meier Curves



Early/Diminishing Effect



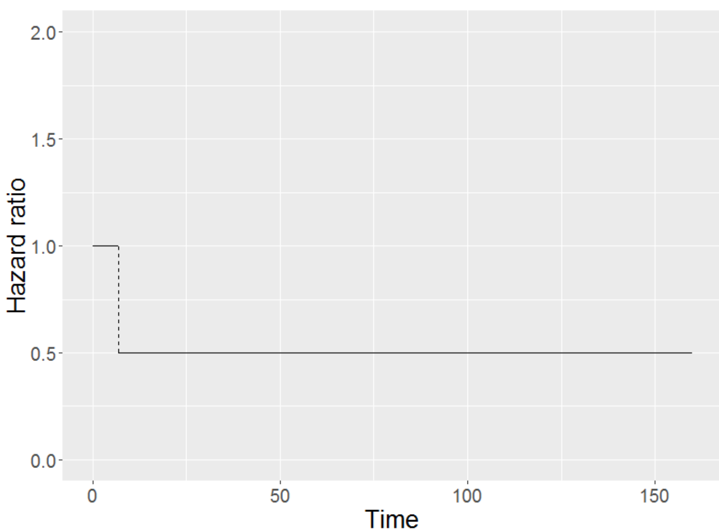


Proprietary

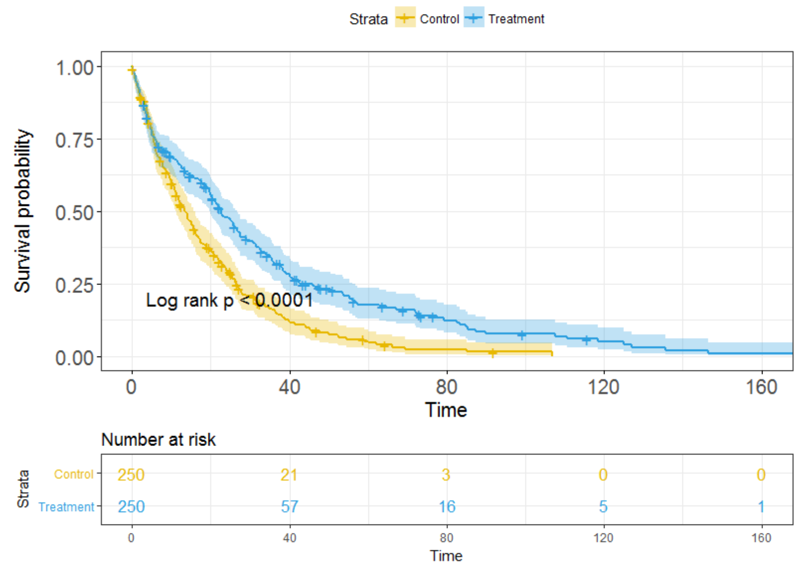
Different Scenarios (non-crossing hazards)

Late/Delayed Effect

Hazard Ratio

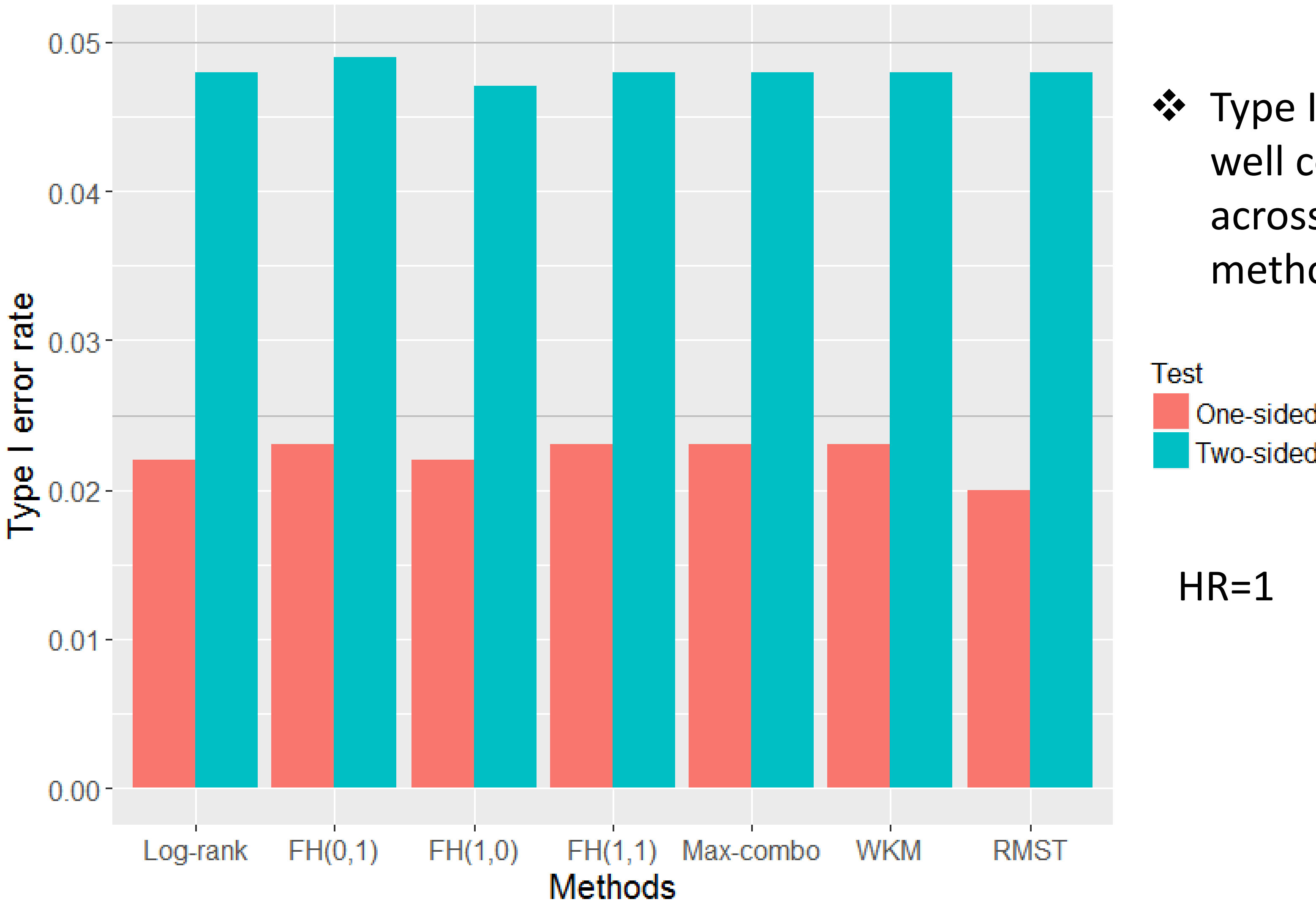


Kaplan-Meier Curves



Proprietary

Comparison of Methods - Type I Error



- Type I error is well controlled across different methods.

$\alpha = 0.025$

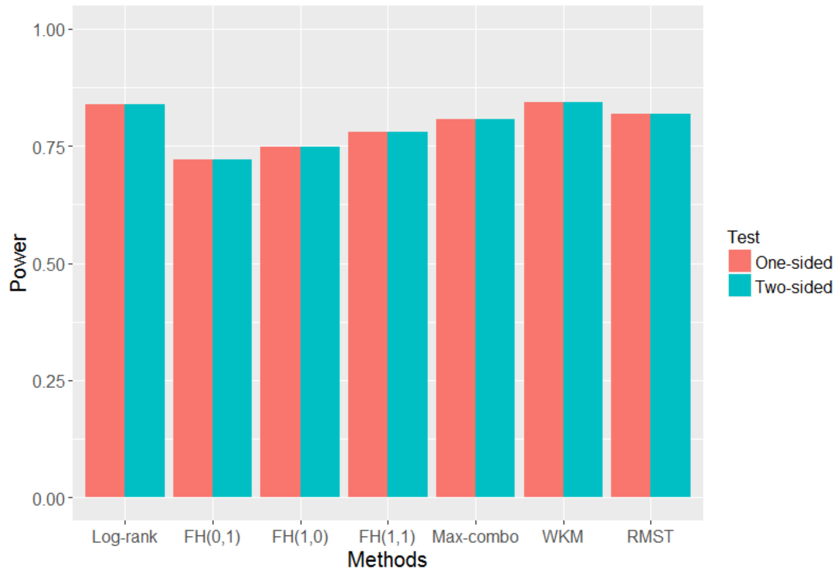
$\alpha = 0.05$

HR=1

Proprietary

Comparison of Methods under non-crossing hazards-Power

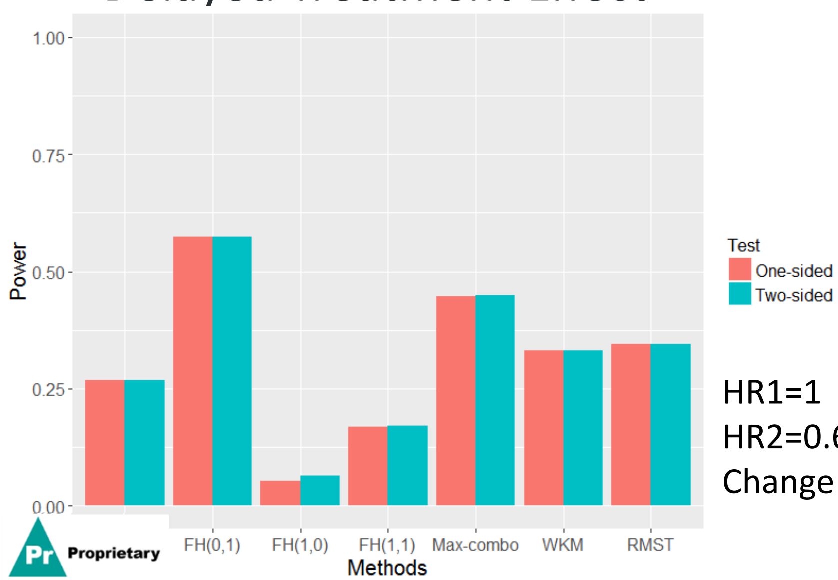
Proportional Hazards



Early Treatment Effect

HR1=0.75
HR2=1
Change pt at 24

Delayed Treatment Effect

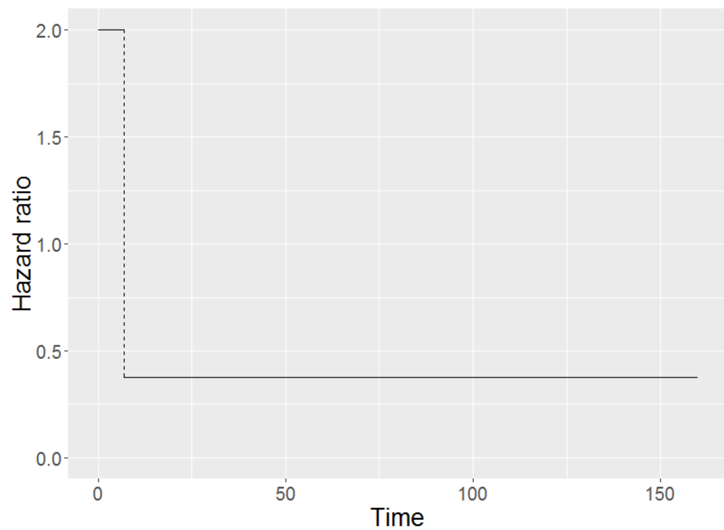


HR1=1
HR2=0.6
Change pt at 7

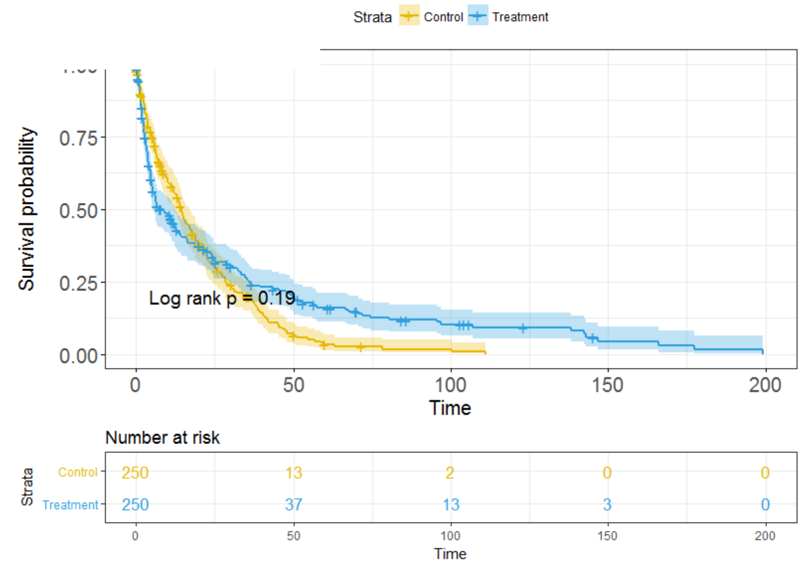
- Max-combo is robust to PH, and early, late effect scenarios of NPH examined.
- WKM less powerful for delayed effect
- One or two sided testing gives similar power.

Crossing Hazards Scenario 1

Hazard Ratio



K-M plot



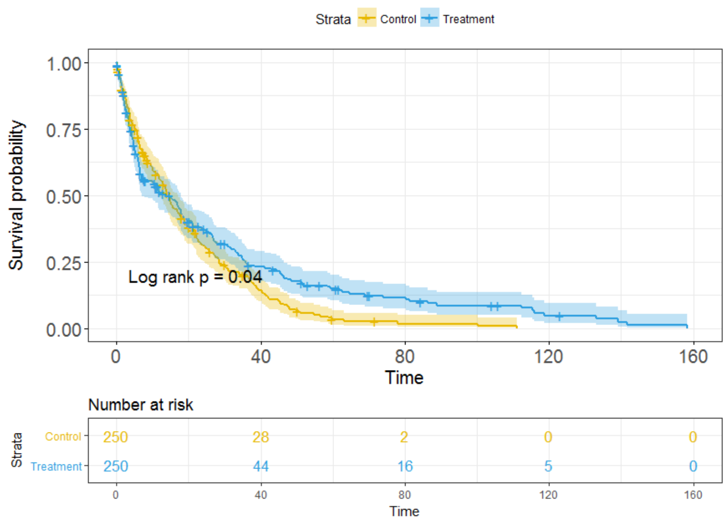
Power

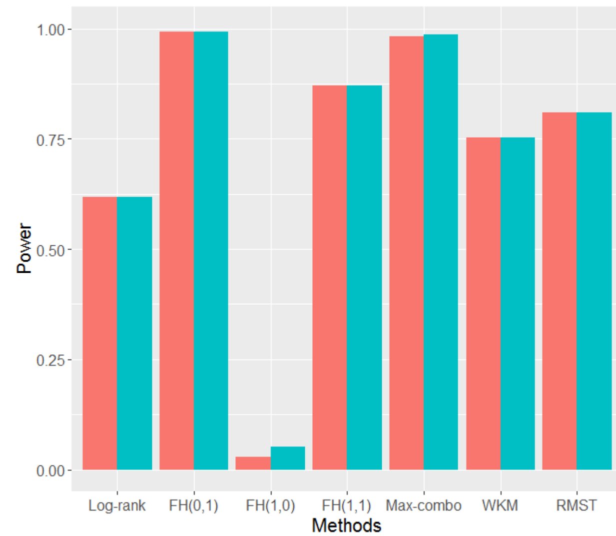
HR1 = 2
HR2 = 0.375
Change pt at 7

Proprietary

Power – Varying Crossing Scenarios

Crossing Hazards 2

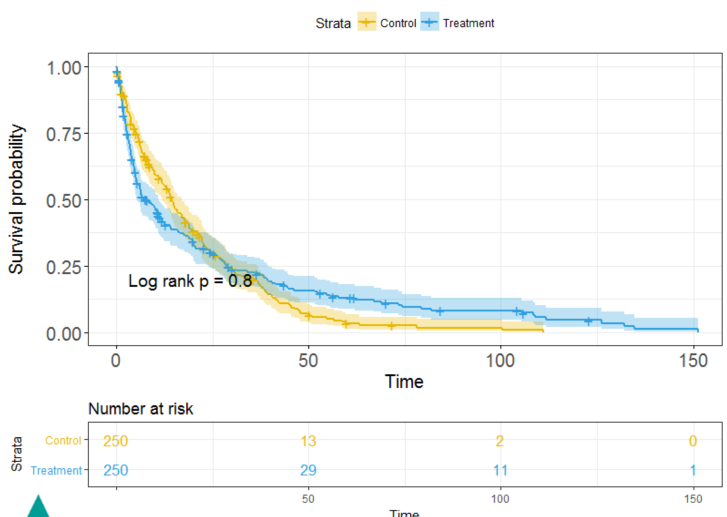


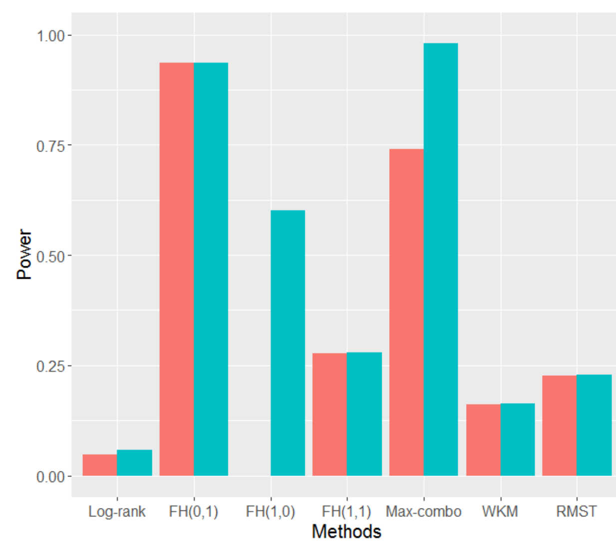


HR1 = 1.5
HR2 = 0.5
Change pt at 7

- One-sided testing gives lower power compared to two-sided testing.

Crossing Hazards 3

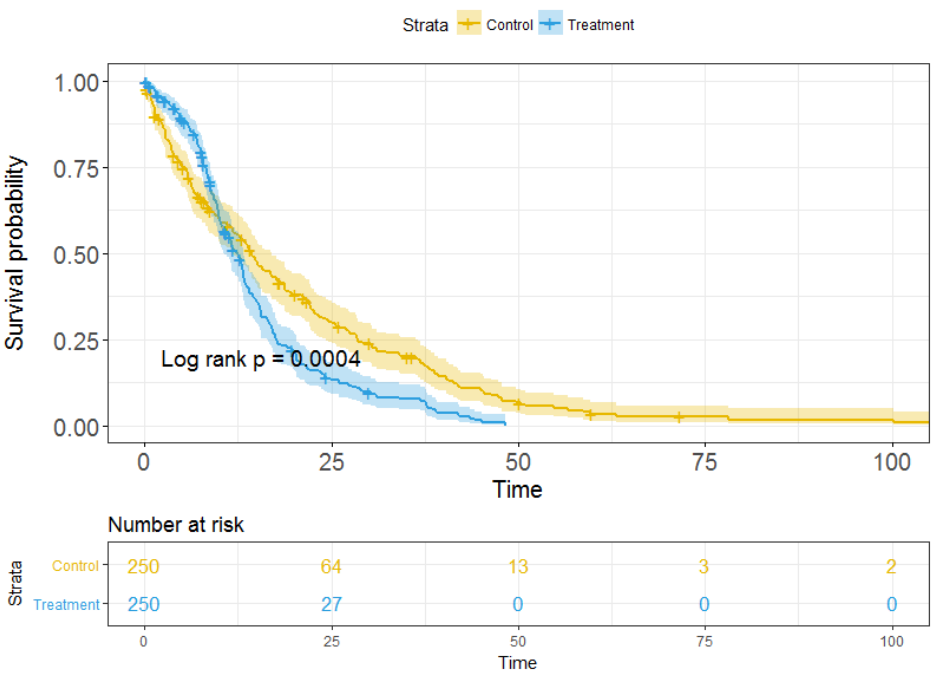




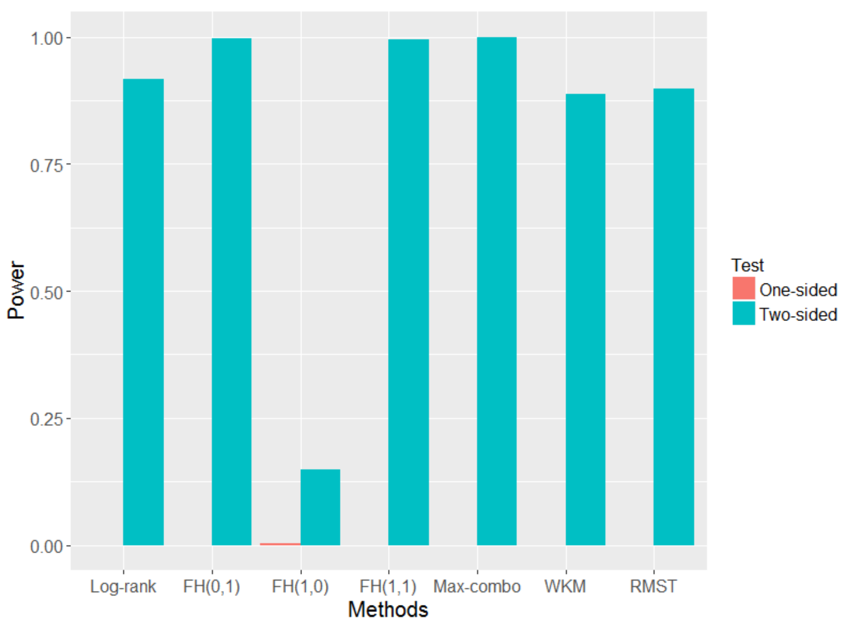
HR1 = 2
HR2 = 0.5
Change pt at 7

Power – Treatment Effect Testing (Cont'd)

Crossing Hazards 4

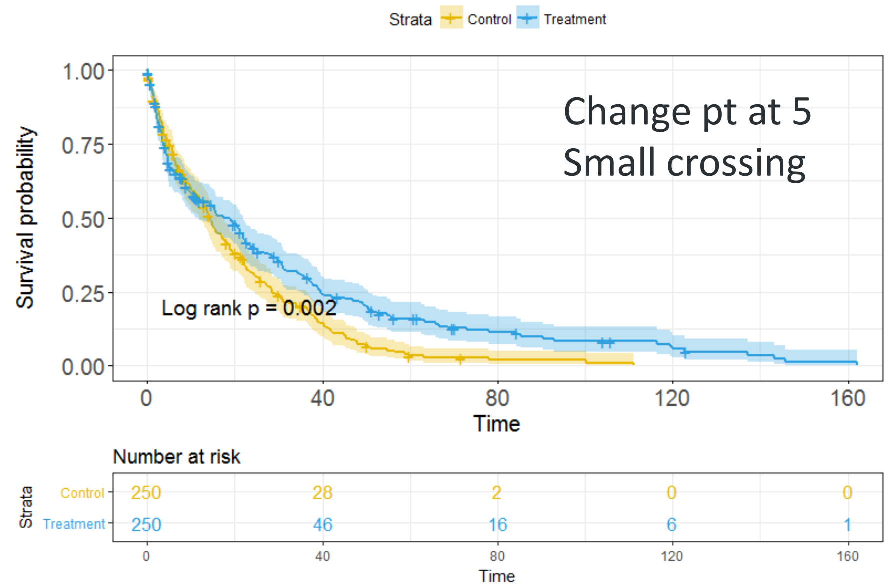


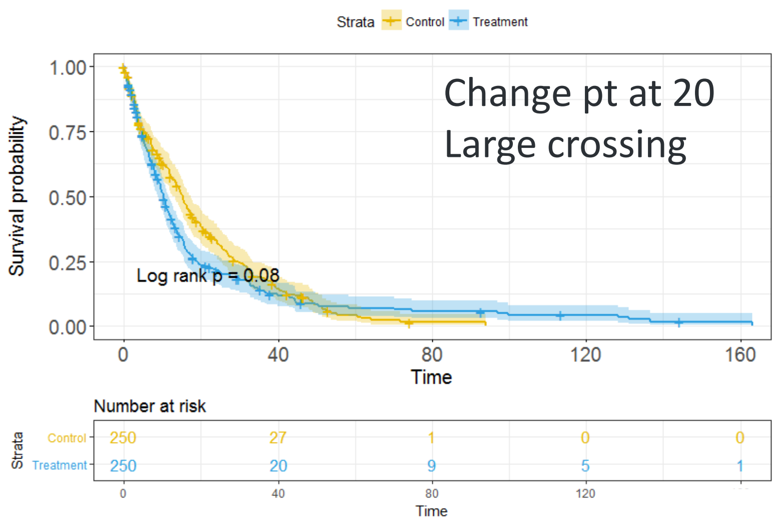
HR1 = 0.5
HR2 = 2
Change pt at 7

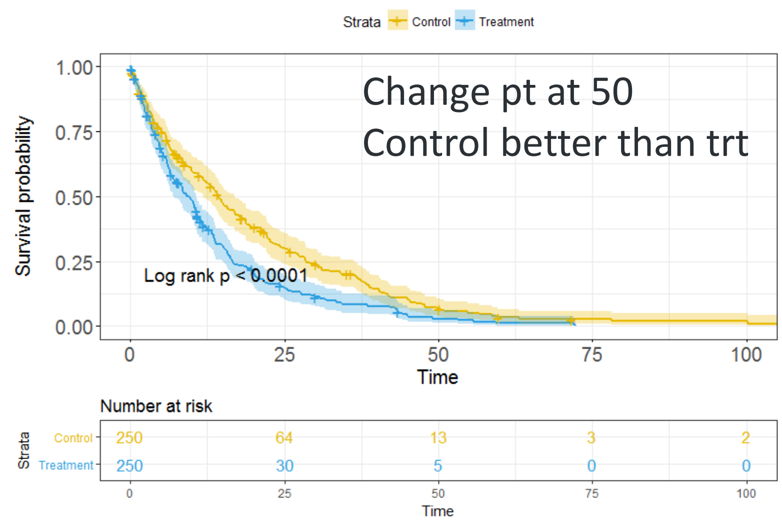


Proprietary

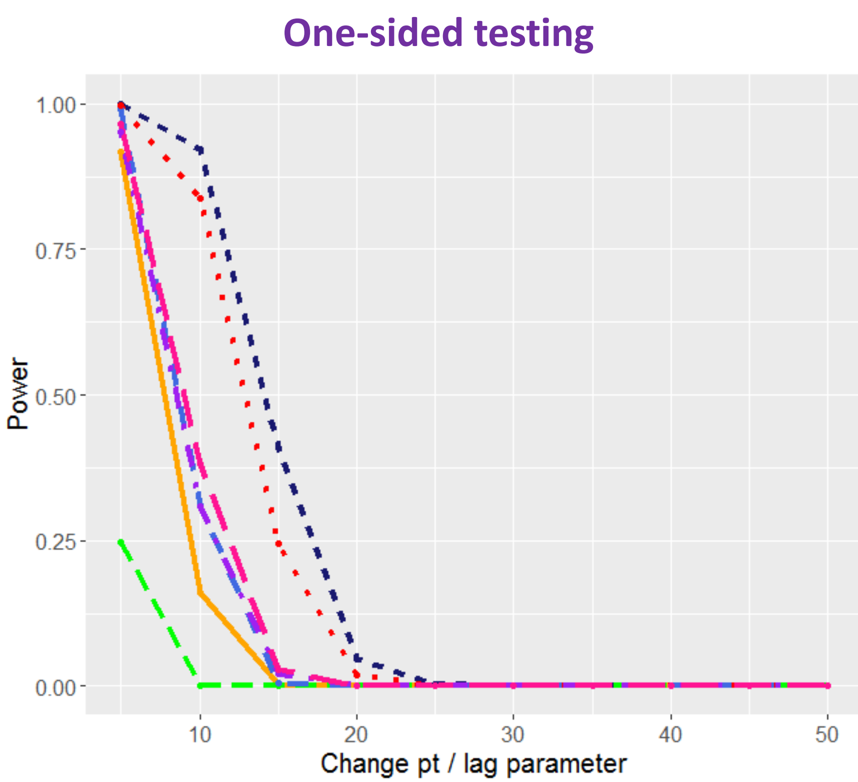
Impact of Change Point Location on Power

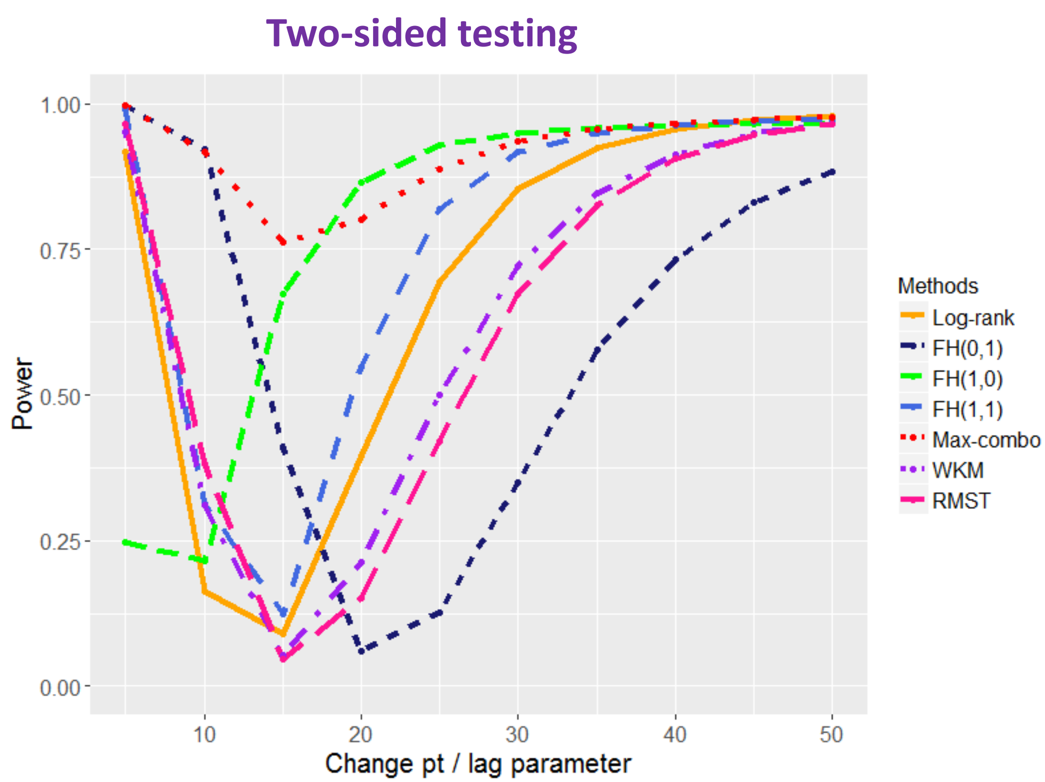






Power is impacted by the location of change points





- Power decreases rapidly as change point moves to later time.
- Power decreases first, then increases as change pt moves to later time.

Proprietary

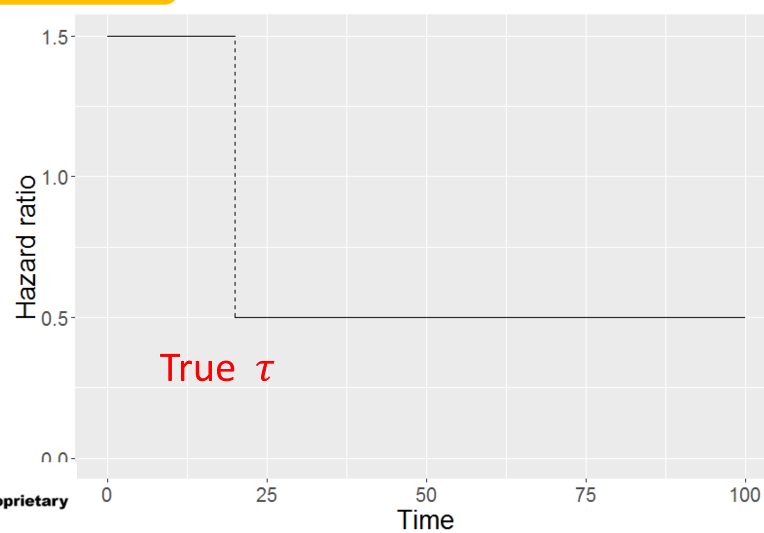
Cox Model with Change Point Model

Treatment Effect Estimation

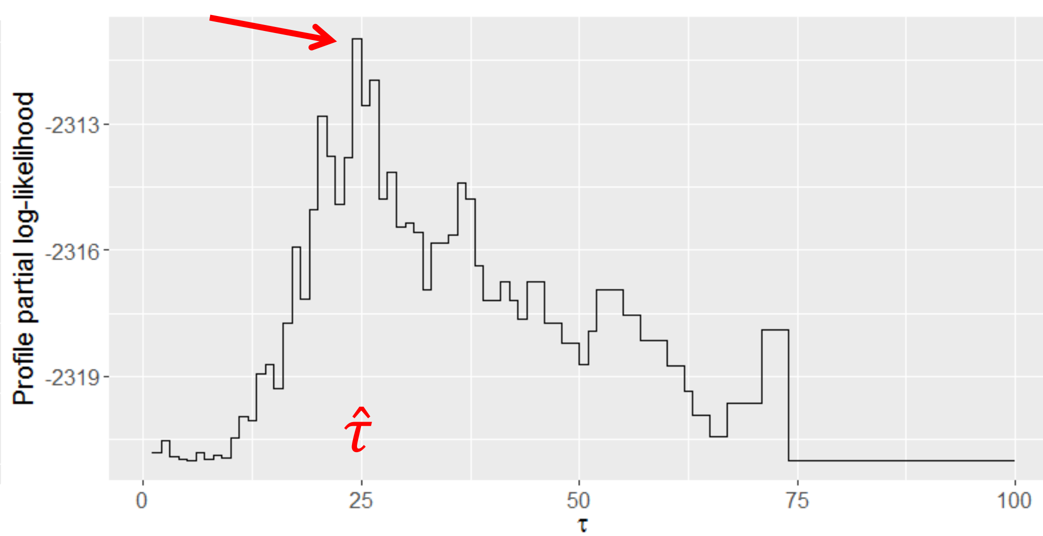
- **Cox PH model with singe change point:** R fct/SAS macro
- **Details**
  - $\lambda(t|Z,\hat{\tau}) = \lambda_0(t) \cdot \exp\left(\beta_1^T Z \cdot 1_{[t\leq\hat{\tau}]} + \beta_2^T Z \cdot 1_{[t>\hat{\tau}]}\right)$
  - Z denotes trt arm (1: experimental; 0: control)
  - τ denotes the change point location (or lag parameter)
    - ❖ $\hat{\tau}$ is estimated through maximizing profile partial likelihood [Liang et al., 1990]

Example

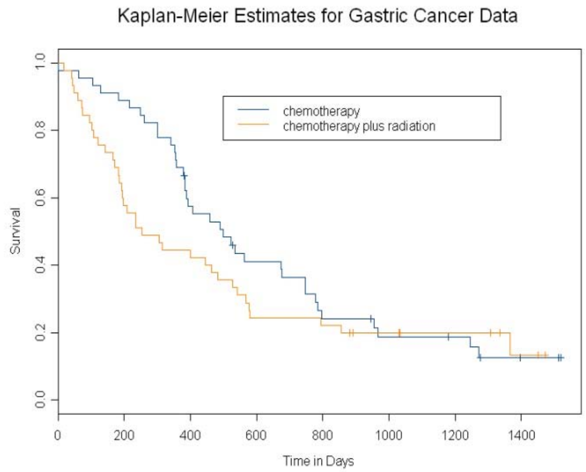
Hazard Ratio

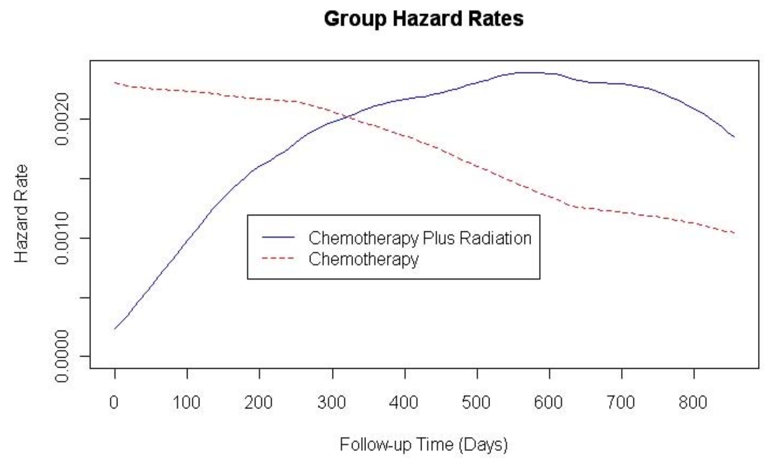


Profile Partial Likelihood wrt τ



Illustrative Example II: Hess (1994)





Over all HR= 1.30 (log rank p-Value 0.630)

Scenario of change point	Cox PH model with single change point		
	Change point (days)	HR1	HR2
Estimated location*	254	4.14	0.62
Location fixed at median of all event times	355	2.77	0.61
Location fixed at median of all observation times	398	1.77	0.83

*change point locations was searched at 0.5 increments, i.e. 0.5, 1, 1.5 etc.

Summary

- Challenging to find one *optimal* analytical method under varying scenarios.
- All methods have their pros and cons

For treatment effect testing under quantitative interaction (no-crossing hazards)

- Max-combo method appears to be robust to different scenarios of NPH examined
  - Requires clinical justification of weight functions
- The G-rho-gamma family of weighted log-rank tests with proper choice of weights have good performance
  - Incorrect weight choice adversely impacts performance
- The weighted Kaplan-Meier test has good performance and is robust for early treatment effect
  - Weights are data driven and do not require pre-specification
- One and two-sided tests give almost same power

For treatment effect testing under qualitative interaction (crossing hazards)

- Most methods lost power under qualitative interaction
  - p-Value may be hard to interpret
  - interpretation of results require visual inspection of data for further interpretation
- one-sided testing gives lower power compared to two-sided testing in most scenarios; one sided test is more appropriate to examine treatment benefit

For treatment effect estimation

- One summary statistics (e.g., HR from Cox PH) may not be sufficient.
  - Cox PH model with change point(s) may serve as an alternative method for NPH especially crossing hazards.
- More work needs to be done...

Selected References

1. Anderson KM, A Non proportional Hazards Weibull Accelerated Failure Time Model. Biometrics, 1991;47:281-288
2. Fleming, T. R. and Harrington, D. P., Counting Process and Survival Analysis. New York, John Wiley and Sons. 1991.
3. Fleming, T. R, Harrington, D. P. and O'Sullivan, M., Supremum Versions of the Logrank and Generalized Wilcoxon Statistics, Journal of the American Statistical Association, 82, (1987): 312-320.
4. **Public Workshop: Oncology Clinical Trials in the Presence of Non-Proportional Hazards, The Duke-Margolis Center for Health Policy, Feb. 2018**
5. Gill, Richard D. "Censoring and stochastic integrals." *Statistica Neerlandica* 34.2 (1980): 124-124.
6. Hess, K., "Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions", Statistics in Medicine, 13:1045-1062, 1994.
7. Kosorok, Michael R., and Chin-Yu Lin. "The versatility of function-indexed weighted log-rank statistics." *Journal of the American Statistical Association* 94.445 (1999): 320-332.
8. Karrison, Theodore G. "Versatile tests for comparing survival curves based on weighted log-rank statistics." *Stata Journal*16.3 (2016): 678-690.
9. Margaret Sullivan Pepe and Thomas R. Fleming, Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data, Biometrics, Vol. 45, No. 2 (Jun., 1989), pp. 497-507
10. Kung-Yee Liang, Steven G. Self and Xinhua Liu, The Cox Proportional Hazards Model with Change Point: An Epidemiologic Application, Biometrics, Vol. 46, No. 3 (Sep., 1990), pp. 783-793
11. Margaret Sullivan Pepe and Thomas R. Fleming, Weighted Kaplan-Meier Statistics: Large Sample and Optimality Considerations, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 53, No. 2(1991), pp. 341-352
12. Mehrotra, D., Su, S., Li, X. An efficient alternative to the stratified Cox model analysis. Statistics in Medicine, 2012*;*31:1849–1856.
13. Hajime Uno, Lu Tian, Brian Claggett and L. J.Wei, A versatile test for equality of two survival functions based on weighted differences of Kaplan–Meier curves, Statistics in Medicine.
14. Pepe M. S. and Fleming T.R., Weighted Kaplan–Meier statistics: a class of distance tests for censored survival data, Biometrics 45 (1989), pp. 497–507.
15. Schumacher, M., Two-sample Tests of Cramer-Von Mises and Kolmogorov-Smirnov Type for Randomly Censored Data. International Statistical Review 52 (1984): 263-281.
16. Yang, S. and Prentice, RL. Semiparametric analysis of short term and long term relative risks with two sample survival data. Biometrika. 2005; 92: 1-17
17. Yang, S and Prentice, R, Improved Logrank-Type Tests for Survival Data Using Adaptive Weights, Biometrics 66(2010): 30-38
18. Tian, Lu; Zhao, Lihui; and Wei, LJ. "On the Restricted Mean Event Time in Survival Analysis." (February 2013). Harvard University Biostatistics Working Paper Series. Working Paper 156. http://biostats.bepress.com/harvardbiostat/paper156

Thank You!