Analysis Methods for Non-Proportional Hazards

Satrajit Roychoudhury , Pfizer Inc.

Duke-Margolis Workshop

Feb 5th, 2018 Washington DC

Acknowledgement

Industry

- Keaven Anderson (Merck & Co.)
- Tianle Hu (Eli Lilly)
- Larry Leon (Roche)
- Pralay Mukhopadhyay (Astrazeneca)
- Honglu Liu (Eli Lilly)

FDA

- Sirisha Musthi (CDER, FDA)

Disclaimer

- Satrajit Roychoudhury is an employee of Pfizer Inc.
- The views and opinions expressed herein are my own and cannot and should not necessarily be construed to represent those of Pfizer Inc. or its affiliates.

Non-Proportional Hazards (NPH): What Does It Mean?

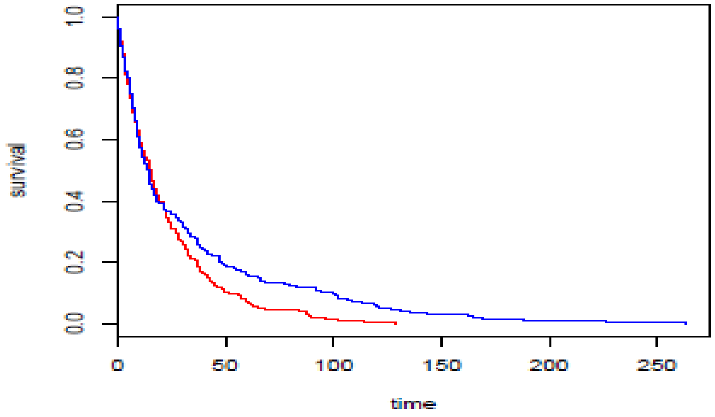
- Most popular methods in randomized clinical trial:
  - **Kaplan-Meier (KM): describe** probability of survival over time
  - **log-rank test (LRT)**: **detect** difference in treatment effect
  - **Cox regression (CR)**: **summarize** the treatment effect
- Log-rank p-value, hazard ratio, and naive median are the standard metrics of reporting
- Are they good summary measures when the treatment effect is not constant over time? : **NPH problem**
  - For example, recent immunotherapy development showed evidence of a delayed effect
- How to cope with NPH problem at design and analysis stages?

Log-rank Test and Cox Regression : Fits to All?

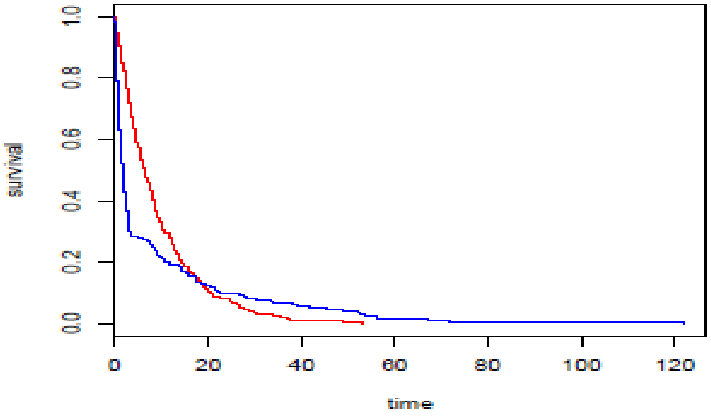
- **LRT :** introduced by Nathan Mantel in 1966
- **CR:** introduced by Sir David R Cox in 1972
- LRT and CR are **closely related**
- LRT is fully nonparametric
  - **most powerful** for proportional hazards (PH)
  - can cause **substantial power loss** if PH assumption does not hold
- Key assumption for CR: **constant** effect over time
  - treatment effect summarized by hazard ratio (HR)
  - problematic if PH assumption violates

Different Types of NPH

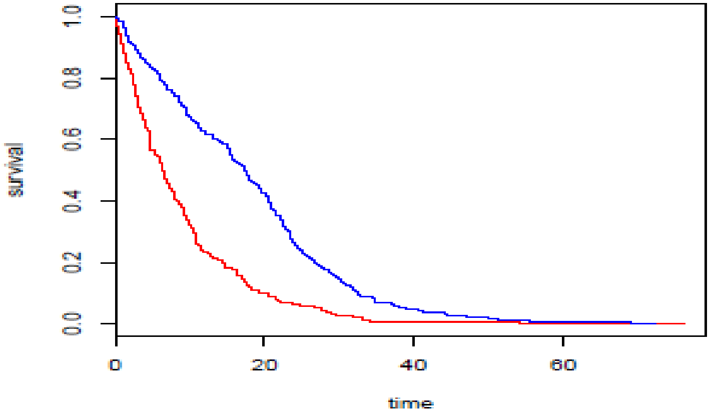
Delayed Treatment Effect



Crossing Hazard



Diminishing Treatment Effect



- Uncertainty related to the type of NPH when trial starts

Analysis and Design Trial with NPH: Key Challenges

- NPH has been discussed extensively in literature
  - alternative methods for hypothesis testing and estimation
- However, application in real life is still rare
- **Main challenge:** NPH type cannot be pre-identified
  - treatment effect profile is unknown at design stage
- **Key questions** for today's forum : in presence of NPH
  - how to plan primary analysis appropriately?
  - How to design a trial?
  - how to efficiently communicate the results with non-statisticians?

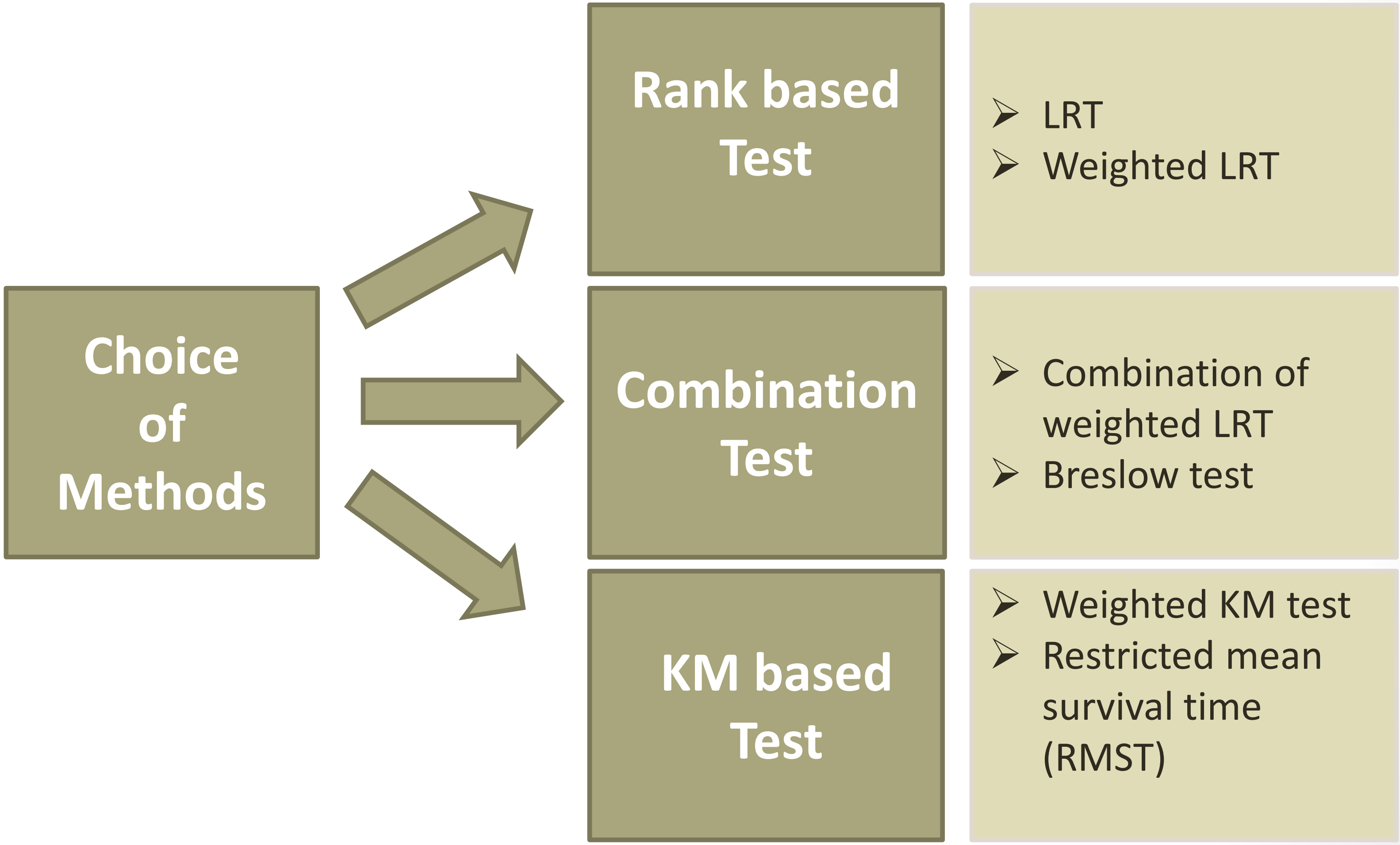
Choice of Primary Analysis in Confirmatory Trials

- Regarding **primary analysis ICH E9** states

  For each clinical trial contributing to a marketing application, all important details of its design and conduct and the principal features of its proposed statistical analysis should be clearly specified in a protocol written before the trial begins. The extent to which the procedures in the protocol are followed and the primary analysis is planned a priori will contribute to the degree of confidence in the final results and conclusions of the trial.

- Specifying primary analysis when NPH is expected**: need robust statistical method** to handle
  - possibility of different types of NPH
  - possibility of different specifications (e.g. lag time for treatment effect)

Choice of Primary Analysis



Choice of Methods

Rank based Test

- LRT
- Weighted LRT

Combination Test

- Combination of weighted LRT
- Breslow test

KM based Test

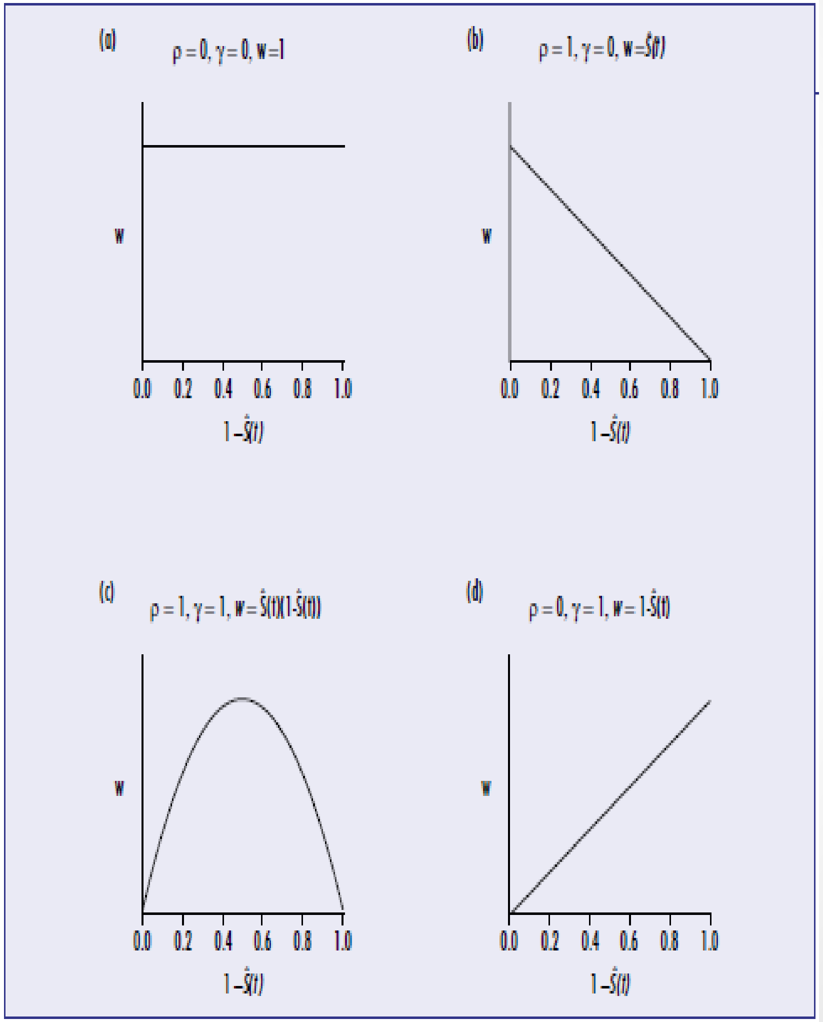
- Weighted KM test
- Restricted mean survival time (RMST)

Weighted Log-rank Test

- Fleming and Harrington proposed a class of weighted log-rank test (FH) based on the $G^{\rho,\gamma}$ family
- Assign weight to events

$W_n(t)= (S_n(t))^{\rho}(1- S_n(t))^{\gamma}$

- Values of ρ and γ implies
  - $\rho > 0, \gamma = 0$: early difference
  - $\rho = 0, \gamma > 0$: late difference
  - $\rho > 0, \gamma > 0$: mid difference
  - $\rho = 0, \gamma = 0$: log-rank test



Combination Test

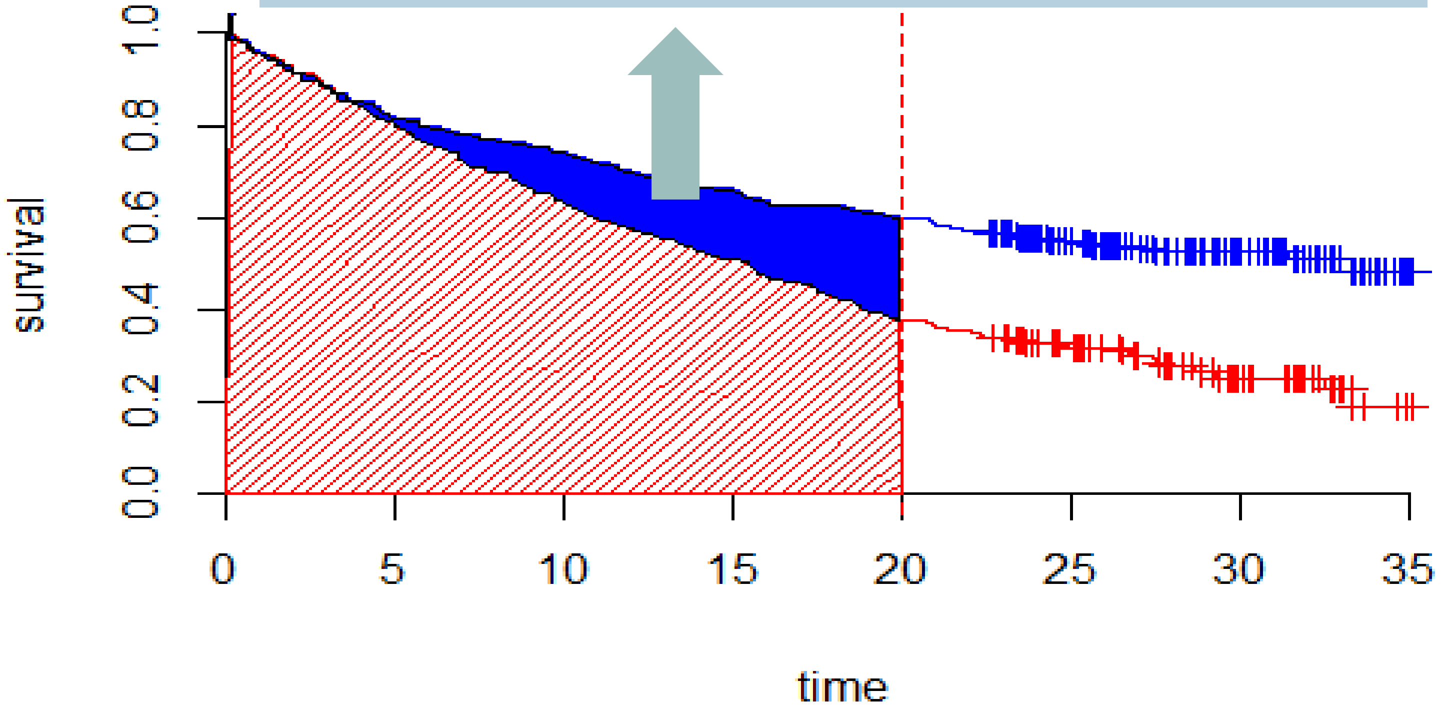
- Major difficulty for FH LRT:
  - specification of ρ and γ parameter: mis-specification may imply a loss of power
- Possible alternative : **Combination test**
  - handles simultaneously a range of NPH types
  - choose the appropriate weight in “adaptive” fashion
- Similar concepts are explored by
  - **Yang and Prentice 2010**: *Adaptively Weighted log-rank Test*
  - Garès et. al. 2017: maximal statistics over FH(0,γ)
  - **Karrison 2016**: *Versatile tests*

Combination of FH Log-rank Test (Max-Combo)

- We have considered two combinations
  - **combination of $G^{0,0}$ and $G^{0,1}$:** ***Combo 1***
  - **combination of $G^{0,0}$, $G^{0,1}$, $G^{1,1}$, $G^{1,0}$:** ***Combo 2***
- **Max-Combo test :** largest of the absolute value of the test statistics
- *"Adaptive"* procedure involving selection of best test statistics: **requires multiplicity correction**
  - Bonferroni-Holmes adjustment (conservative)
  - adjustment using the joint asymptotic distribution of the FH log-rank test statistics (**recommended**)
- Can be pre-specified easily at protocol stage **: satisfies ICH E9 condition**

Kaplan-Meier Based Tests

- Treatment effect (Difference scale) at month 20
- **KM based test** are based on the difference/ratio between two KM curves



Kaplan-Meier Based Tests

- **Weighted Kaplan-Meier test: (Pepe and Fleming, 1989, 1991)**
  - weighted difference of area under KM curves up to **a specified cut-off**
  - weights are based on KM estimate of censoring
  - need to specify **the cut-off**: can be affected by censoring
- **Restricted mean survival time (RMST) (Uno *et al* 2014)**
  - area under the KM plot prior to specific time-point: can be easily interpreted as "life expectance"
  - treatment effect: difference or ratio of RMST
  - need to specify **the cut-off**: can be affected by censoring

Other Methods

- **Piecewise log-rank test (Xu. *et al* 2016)**
  - piecewise weighted log-rank test within specified time intervals
  - optimal when weights for earlier events are zero
  - *power/type-I error* ***greatly affected if intervals are incorrectly specified***
- **Other combination tests :**
  - **Breslow et. al. 1984:** combination of log-rank test and test of acceleration
  - **Logan 2008:** combination of log-rank test and milestone survival, it suffers similar problem as other KM based tests
- In the next talk the simulation study results will be presented

Reporting Treatment Effect

- When NPH is present: HR depends on time
  - HR or average HR as a single number is less useful
  - *what statistics to be reported to quantify treatment effect?*
  - *how to appropriately pre-specify to meet ICH E9?*
- A **sequential approach (Royston and Parmer 2010)**
  - **First step:** perform Max-combo test to conclude about the "Null" hypothesis (no treatment effect)
  - **Second step:** regardless of results in step 1, gather evidence of NPH, possible options
    - Grambsch–Therneau test for PH
    - other graphic diagnostics for confirming PH
  - **Third step**: choose treatment effect summary based on step 2- *treatment effect estimate beyond test statistics*

Choice of Treatment Effect Summary

- If PH assumption is reasonable
  - **HR from Cox regression (CR)** and corresponding 95% confidence interval (CI)
  - secondary analysis**:** average HR from weighted CR and 95% confidence interval (weight chosen by Max-combo)
- If there is evidence of NPH, the possible metrics
  - **ordinary/average HR** with 95% CI (Max-combo estimate)
  - **difference in RMST at t^*:** gain in *life expectancy* at clinically relevant time point t^* (pre-specified)
  - **difference in milestone survival at t^*:** gain in chance of survival at clinically relevant time point t^* (pre-specified)
  - secondary analysis**:** piecewise HR with 95% CI
- In **session III**, case studies will elaborate this approach

Conclusion

- NPH team looked into different possible methodologies
- Max-combo looks a promising approach
  - allows possibility for different NPH type
  - provides robustness under model mis-specification
- In presence of NPH a single measure is less useful
  - a sequential approach can be useful
- Team has included all the procedures in a R package "**nphsim**" : freely available from *github*
- In next talks, the team members will present simulation results and case studies

References

1. Breslow, N *et. al.* (1984). A two sample censored data rank test for acceleration. *Biometrics,* 40: 1042–1069
2. Cox DR (1972). Regression-Models and Life-Tables. *Journal of the Royal Statistical Society*. B (Methodological), 34 (2): 187–220
3. Fleming TR, Harrington DP (1991) Counting Processes and Survival Analysis. John Wiley & Sons: New York
4. Garès V *et. al*. (2017). On the Fleming–Harrington test for late effects in prevention randomized controlled trials. Journal of Statistical Theory and Practice, 11(3): 418-435
5. ICH E9 http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E9/Step4/E9_Guideline.pdf
6. Karrison TG (2016). Versatile tests for comparing survival curves based on weighted log-rank statistics. Stata Journal, 16(3): 678-690

References

7. Lin DY, Wei LJ (1989). The robust inference for the Cox proportional hazards model. Journal of the American Statistical Association, 84:1074–1078
8. Logan, BR *et. al.* (2008), Comparing Treatments in the Presence of Crossing Survival Curves: An Application to Bone Marrow Transplantation. Biometrics, 64: 733–740
9. Mantel N (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50 (3): 163–70
10. Pepe MS, Fleming TR (1989). Weighted kaplan-meier statistics: A class of distance tests for censored survival data. Biometrics, 45(2):497–507
11. Pepe MS, Fleming TR (1991). Weighted kaplan-meier statistics: Large sample and optimality considerations. Journal of the Royal Statistical Society Series B (Methodological), 52:341–352

References

11. Royston P and Parmar MK (2010). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. Statistics in Medicine, 21(15): 2175-2197
12. Uno H *et. al.* (2014). Moving beyond the hazard ratio in quantifying the between-group difference in survival analysis. J. Clinical. Oncol., 32(22): 2380-2385
13. Xu Z *et. al.* (2017). Designing therapeutic cancer vaccine trials with delayed treatment effect. Statistics in medicine, 36(4):592-605
14. Yang S, Prentice R (2010). Improved Logrank-Type Tests for Survival Data Using Adaptive Weights. *Biometrics*, 66(1): 30-38

Back-up

Max-Combo Test

Let, $Z_1 = \mathbf{G}^{0,0}$, $Z_2 = \mathbf{G}^{0,1}$, $Z_3 = \mathbf{G}^{1,1}$, and $Z_4 = \mathbf{G}^{1,0}$

Max-Combo Test : $\mathbf{Z}_{\mathbf{max}} = \max(|Z_1|, |Z_2|, |Z_3|, |Z_4|)$

Under $\mathbf{H_0}$, $(Z_1, Z_2, Z_3, Z_4) \sim N_4(\mathbf{0}, \mathbf{\Sigma})$ (Karrison et. al 2016)

$\mathbf{\Sigma} = (\sigma_{ij})_{4X4}$; $\sigma_{ij} = \mathrm{cov}(\mathbf{G}^{a,b}, \mathbf{G}^{c,d}) = V(\mathbf{G}^{a+c/2,b+d/2})$: a,b,c,d = 0 or 1

The p-value for $\mathbf{Z}_{\mathbf{max}}$ can be derived by integrating under the multi-variate normal density

Average Hazard Ratio

- Average hazard ratio (AHR) represents the “average effect” of treatment over the course of the trial
- Associated estimator of Max-Combo test: AHR using weighted cox regression (WCR)
- Choosing weight (ρ,γ) that provides maximal test statistics
  - variance: robust estimate proposed by Lin and Wei 1989
  - point estimate and 95% confidence interval
  - multiplicity adjusted confidence internal using null distribution of Max-Combo
- However, the WCR under non-proportional hazards lack intuitive simplicity