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**

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

• 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

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

\( \hat{S} \) 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:

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

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

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
    \[ FH(\rho, \gamma) = S(t)^\rho \cdot (1 - S(t))^{\gamma} \]
  • 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
    \[ \hat{w}_c(t) = \frac{\hat{C}_1(t)\hat{C}_2(t)}{\hat{p}_1\hat{C}_1(t) + \hat{p}_2\hat{C}_2(t)}, \]
    where \( \hat{C}_i(t) \) is prob of not being censored before time \( t \)
    (i.e., censoring survival function)
    weights monotonically decreasing with time
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 $FH(\rho, \gamma)$ 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:
  \[ Z_{\text{max}} = \max(|Z_1|, |Z_2|, |Z_3|, |Z_4|) \]
- Under $H_0$, $(Z_1, Z_2, Z_3, Z_4) \Rightarrow MVN_4(0, \Sigma)$
  - $\Sigma = (\sigma_{ij})_{4x4}$, where
    \[ \sigma_{ij} = \frac{n_1+n_2}{n_1n_2} \int_0^{\infty} K_i(t) K_j(t) \frac{Y_1(t)Y_2(t)}{V_1(t)+V_2(t)} \left( 1 - \frac{\Delta N_1(t)+\Delta N_2(t)-1}{V_1(t)+V_2(t)-1} \right) \left[ \frac{d(N_1(t)+N_2(t))}{V_1(t)+V_2(t)} \right] \]
  - 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
Different Scenarios (non-crossing hazards)

Late/Delayed Effect

Hazard Ratio

Kaplan-Meier Curves

Log rank p < 0.0001
Comparison of Methods - Type I Error

- Type I error is well controlled across different methods.

HR = 1

<table>
<thead>
<tr>
<th>Test</th>
<th>One-sided</th>
<th>Two-sided</th>
</tr>
</thead>
<tbody>
<tr>
<td>α</td>
<td></td>
<td></td>
</tr>
<tr>
<td>α = 0.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>α = 0.05</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Diagram showing type I error rates for various methods]
Comparison of Methods under non-crossing hazards - Power

- **Proportional Hazards**
  - **Early Treatment Effect**
  - **Delayed Treatment Effect**

- **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.

- **HR1** = 0.75
  - HR2 = 1
  - Change pt at 24

- **HR1** = 1
  - HR2 = 0.6
  - Change pt at 7
Crossing Hazards Scenario 1

Hazard Ratio

K-M plot

Power

HR1 = 2
HR2 = 0.375
Change pt at 7
Power – Varying Crossing Scenarios

Crossing Hazards 2

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

Crossing Hazards 3

- HR1 = 1.5
- HR2 = 0.5
- Change pt at 7

- HR1 = 2
- HR2 = 0.5
- Change pt at 7
Power – Treatment Effect Testing (Cont’d)

Crossing Hazards 4

Log rank $p = 0.0004$

HR1 = 0.5
HR2 = 2
Change pt at 7
Impact of Change Point Location on Power

Change pt at 5
Small crossing

Log rank p = 0.002

Change pt at 20
Large crossing

Log rank p = 0.08

Change pt at 50
Control better than trt

Log rank p < 0.0001
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.
Cox Model with Change Point Model
Treatment Effect Estimation

• Cox PH model with single change point: R fct/SAS macro
• Details
  – \( \lambda(t|Z, \hat{\tau}) = \lambda_0(t) \cdot \exp(\beta_1^T Z \cdot 1_{[t \leq \hat{\tau}]} + \beta_2^T Z \cdot 1_{[t > \hat{\tau}]} ) \)
  – \( Z \) denotes trt arm (1: experimental; 0: control)
  – \( \tau \) denotes the change point location (or lag parameter)
    ✷ \( \hat{\tau} \) is estimated through maximizing profile partial likelihood [Liang et al., 1990]
Illustrative Example II: Hess (1994)

**Kaplan-Meier Estimates for Gastric Cancer Data**

**Group Hazard Rates**

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

<table>
<thead>
<tr>
<th>Scenario of change point</th>
<th>Cox PH model with single change point</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Change point (days)</td>
</tr>
<tr>
<td>Estimated location*</td>
<td>254</td>
</tr>
<tr>
<td>Location fixed at median of all event times</td>
<td>355</td>
</tr>
<tr>
<td>Location fixed at median of all observation times</td>
<td>398</td>
</tr>
</tbody>
</table>

*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

Thank You!