Weighted Log-rank Tests and Weighted Cox Models in Non-Proportional Hazards

Ray Lin
Genentech/Roche

San Francisco Bay Area Chapter of American Statistical Association Monthly Seminar
May 21, 2018
Agenda

- Non-proportional Hazards (NPH) in time-to-event endpoints
- Impact on analyses and clinical interpretation
- Cross-Pharma Working Group
- Alternative analysis methods
  - Weighted log-rank test
    - Simulation Study #1
  - Weighted Cox model
    - Simulation Study #2
    - Example on a real study
    - Simulation Study #3
- Ongoing efforts
  - Max-Combo test
  - Alternative study design and analysis
Non-Proportional Hazards (NPH)

- NPH = hazard ratio (treatment effect) changes over time
  - Delayed treatment effect (late separation of curves)
    - Cancer immunotherapies (due to its MOA)
  - Diminishing effect (merging curves)
    - Overall survival confounded by subsequent therapies
  - Crossing survival curves
    - Detrimental during a certain period of time

Delayed Effect (OAK)

Diminishing Effect (BOLERO II)


Impact on Analysis

1. Log-rank test power is reduced by 30% for a 3-month delay (with 12-month median control)

2. Hazard ratio estimate is biased (Cox model)
   - Proportional hazards (constant hazard ratio) assumption is violated
   - Treatment effect is diluted
     - Cox model averages out the treatment effect across all time points

\[
HR = \begin{cases} 
1 & \text{for } t \leq t^0 \\
0.7 & \text{for } t > t^0 
\end{cases}
\]
1. Clinical interpretation and decision making
   - Hazard ratio = risk reduction
   - Potentially misleading if only reporting a constant effect over time based on the standard Cox model
     - E.g., a patient who cannot stay on treatment for a sufficient time may be unlikely to get benefits

2. Evaluation for health economics
Cross-Pharma Working Group on NPH

- **Objectives:**
  - To address the issue of NPH for design, analysis and interpretation
    - Across Oncology, ImmunoOncology
    - Focus on Phase III / regulatory trial setting

- **Workstreams**
  - Endpoint
  - Design and Simulation
  - Analysis

- **Milestone and goal:**
  - Duke Margolis workshop Feb. 5, 2018 with FDA and EMA participants
  - Manuscripts 2018
  - Guidance and White Papers 2018

- **Members**
  - AZ, BMS, Merck, B&I, Novartis, Lilly, Abbvie, Roche, Bayer, Janssen, Takeda, Amgen, Pfizer, GSK, Celgene, and FDA
1. Hypothesis testing
   – Weighted log-rank tests
   – Restricted Mean Survival Time (RMST)
   – Weighted KM tests
   – Testing difference of survival rates at specific time points

2. Treatment effect estimate
   – Hazard ratio estimate based on weighted Cox models
Weighted Log-Rank Tests

- Log-rank test
  \[ Z = \frac{\sum_{j=1}^{J} (O_{1j} - \frac{O_j}{N_j} N_{1j})}{\sqrt{\sum_{j=1}^{J} V_j}} \]

- Weighted Log-rank test
  \[ Z = \frac{\sum_{j=1}^{J} W_j (O_{1j} - \frac{O_j}{N_j} N_{1j})}{\sqrt{\sum_{j=1}^{J} W_j^2 V_j}} \]

- Weight functions
  - Fleming-Harrington weight:
    \[ W(t) = S(t)^{\rho} (1 - S(t))^\gamma \]
    - Log-rank test (0,0)
    - Wilcoxon-Prentice (1,0)
  - Piece-wise constant weight
Simulation Study #1

Design

- **NPH patterns**
  - Delayed effect

- **Weight functions**
  - Fleming-Harrington family

- **Settings**
  - \( N = 400 \) (1:1 randomization)
  - Treatment effect HR: 0.55, 0.7
  - Control median survival: 12 or 24 mos
  - Enrollment: 12 mos with ramp-up or uniform
  - Delay pattern: delayed vs. crossing effect
  - Delay duration: 10% -- 50% of control median
  - Data maturity: event-patient ratio: 60% vs 70%
Simulation Study #1

Summary

- Type-I error is preserved
- Choice of weights in Fleming-Harrington family
  - < 10% delay → LR
  - 10% - 40% → WLR(1,1)
  - > 40% → WLR(0,1)
- Consistent results across
  - Recruitment patterns
  - Median survival
  - Data maturity
  - Treatment effect (HR)

Fleming-Harrington weight family

![Graph showing Fleming-Harrington weight family](image)
1. What is the clinical interpretation of the weights?

2. What is the corresponding treatment effect estimate?

We proposed a hazard ratio time-profile estimate based on “weighted” Cox models
Log-rank statistics is equivalent to the Score statistics in Cox model.

**Cox Model**
- $X$: treatment assignment (tx=1; ctrl=0)
- $\beta$: treatment effect (usually report $\text{exp}(\beta)$ as “hazard ratio”)
- Hazard: $h(t) = h_0(t)\text{exp}[\beta X]$, $h_0(t)$ is baseline hazard
- Survival: $S(t) = \text{exp}[-\int_0^t h(s)ds]$

HR estimated using Cox model (ie, $\text{exp}[\beta]$) corresponds to hypothesis testing using log-rank test.
“Weighted” Cox Model

For a weighted log-rank test with normalized $W(t)$, let a Cox model:

- Hazard: $h(t) = h_0(t) \exp[\beta W(t)X]$
- Score statistics is equivalent to the weighted log-rank
- Partial (%) treatment effect
  - Let $Y = W(t)X$ (a time-dependent covariate)
  - Create $Y$, fit the Cox model and get the coefficient estimate for $Y$ (ie, $\beta$)
- Time-Varying HR (time-profile)
  - $HR(t) = \exp[\beta W(t)]$

Notes:
- $W(t)$ in $[0, 1]$ with maximum 1. Normalization of $W(t)$ will not change the statistics
- Lin (1991) and Sasieni (1993) incorporate weight functions into the score function; the score statistics is equivalent to ours.

Simulation Study #2

Design

- NPH patterns
  - Delayed effect
  - Diminishing effect

- Weight functions
  - Piece-wise linear and exponential tail

- Settings
  - N= 400, 720 (1:1 randomization)
  - Treatment effect HR: 0.68, 0.75
  - Varying
    - Delayed effect (or prolonged effect) duration
    - Effect % during the delayed /prolonged effect duration
  - Control median survival: 12 mos
  - Enrollment: 12 mos with ramp-up
  - Data maturity: event-patient ratio: 70%

Simulation Study #2
Summary

- Type-I error is preserved

- Power
  - The most powerful test if the weight assumption is correct
  - Still more powerful than standard log-rank tests even if mis-specified

- Hazard ratio estimate
  - Unbiased if the assumption is correct
  - Still less biased than standard Cox models even if mis-specified

Example from a Real Study
0-1 Piece-wise Constant Weight

- Piece-wise constant with 0 or 1 weight
- Clear clinical interpretation of the HR estimates
  - HR after $t_0$
  - Risk reduction after $t_0$ among the patients who survive through $t_0$
- Equivalent to landmark analysis
Poplar (GO28753) 2L NSCLC

Effect assumption | P value* | HR Estimate*
--- | --- | ---
Always full effect (Standard Log-rank) | 0.0056 | 0.68 (0.51, 0.89)
Minimal first 3 mo; full effect after 3 mo | 0.0020 | 0.61 (0.45, 0.84)
Minimal first 8 mo; full effect after 8 mo | 0.0006 | 0.50 (0.34, 0.75)

*Unstratified analysis based on data cut Dec 1, 2015
Simulation Study #3
Design

- Scenarios
  - Delayed effect
- Weight function
  - 0-1 weights
- Settings
  - N = 400 (1:1 randomization)
  - Event-patient ratio 70%
  - Treatment effect HR: 0.68
  - Control median survival: 12 mos
  - Varying
    - Delay duration ($t_0$)
    - Adjustment factor ($a_0$)
  - Enrollment: 12 mos with ramp-up
Power and HR Estimate
Varying $t_0$

- **Truth**
  - $a_0 = 0$
  - $t_0 = 3$ mos

- **Model**
  - $a_0 = 0$
  - $t_0 = 0$ to 6 mos

% Effect (Truth) vs Weight (Model)
Power and HR Estimate
Varying $a_0$

- **Truth**
  - $a_0 = 0$ to 100%
  - $t_0 = 3$ mos

- **Model**
  - $a_0 = 0$
  - $t_0 = 3$ mos

Graphical representation showing the relationship between $a_0$ and the effect, as well as the weight model.
Simulation Study #3

Summary

- Type-I error is preserved

- Hazard ratio estimate
  - 0-1 weights is easier to interpret (than Fleming-Harrington weights)
  - Unbiased if the assumption is correct or mis-specified in the conservative direction
  - May biased toward null but still less biased than Cox models if mis-specified
  - Not biased away from null

- Power
  - Most powerful if weight is correctly specified
  - In general more powerful than log-rank
  - May be less powerful than log-rank if weight assumption is far off
Pre-Specification of the Weights

The weight function needs to be pre-specified in order to preserve Type-I error

- Chosen based on prior data or scientific rationales
- Need to evaluate the “loss” if mis-specified (via simulation)
  - In general more powerful than log-rank
  - May be less powerful if weight function is far off

Max-Combo test (work in progress)

- Pre-specifying several weight functions
- Report the one with the max Z score
- P value adjusted for model selection
- HR is tricky to adjust
  - Choose 0-1 weight family and report all corresponding HR estimates
Rationales:
- Based on prior studies and the MOA of the molecule, a delayed treatment effect is expected in this study.
- The duration of the delay is not clear.

Design and analysis plan:
- Propose 0-1 weight functions, assuming the delay duration is
  - Short: 0 month (log-rank)
  - Medium: 3 months
  - Long: 6 months
- Testing: use the Max-Combo test
- Estimate: report the estimates from all weight functions
  - Overall HR (Cox); HR after 3 months; HR after 6 months

Run simulations to characterize these weight functions (gain and loss if the model is correctly specified or mis-specified)
Weighted log-rank tests and Cox models may be used as alternative analysis methods under NPH

- Focus analysis on the time points where the treatment effect is less diluted
- Achieve higher power than standard log-rank test
- Enable reporting of a hazard ratio time-profile
  - Less biased hazard ratio estimate than standard Cox model
  - Potentially more informative description of clinical benefit
Acknowledgements

- Roche/Genentech
  - Larry Leon
  - Zhengrong Li
  - Ina Rhee
  - Jing Yi
  - Ru-Fang Yeh
  - Ben Lyons
  - Gracie Lieberman
  - Yumeng Li
  - Shi Li
  - Na Cui
  - CIT Endpoint Taskforce

- Non-Proportional Hazards Cross-Pharma Working Group
  - Design and simulation subteam
  - Analysis subteam