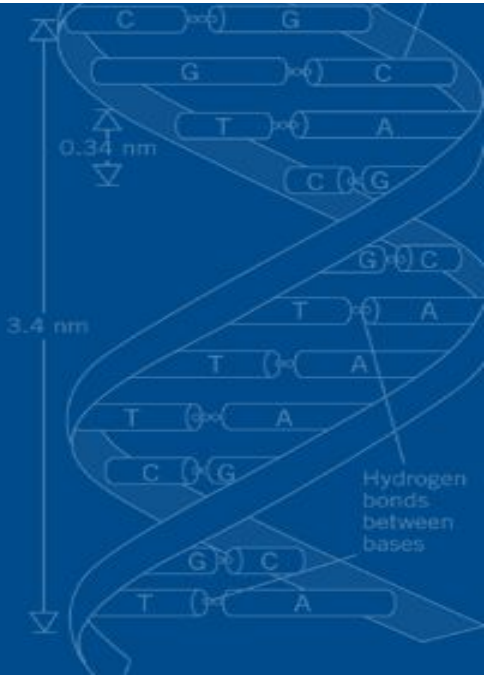
Estimation of treatment effect under non-proportional hazards

Ray Lin

Genentech/Roche

Sept. 24, 2019

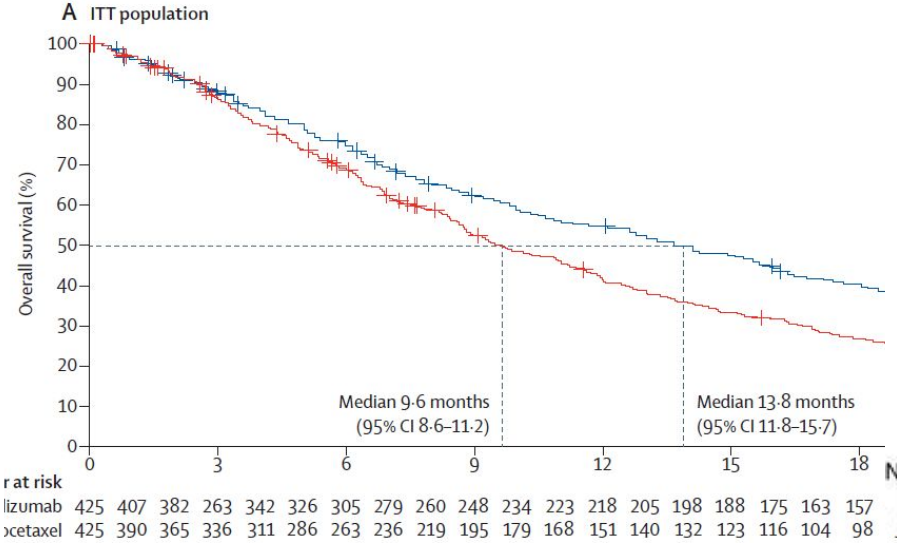


Non-Proportional Hazards (NPH)

NPH = hazard ratio (treatment effect) changes over time

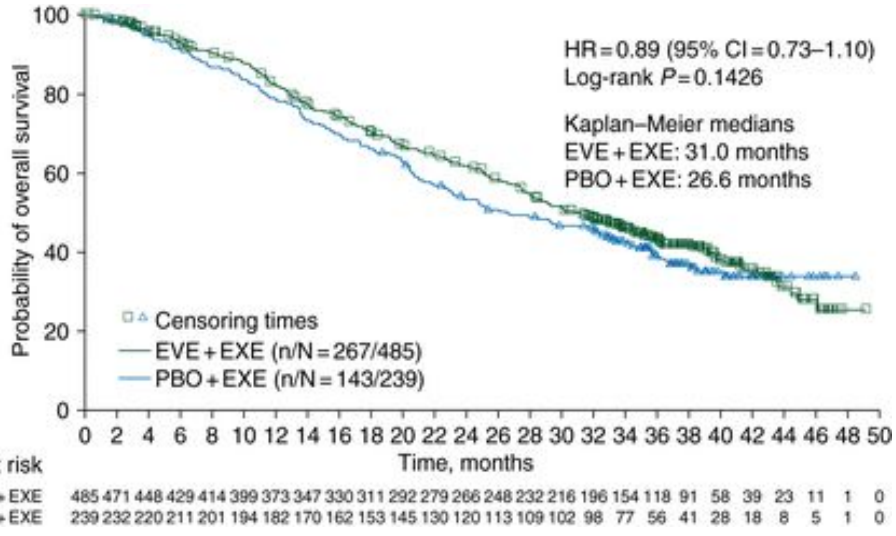
- Delayed treatment effect (late separation of curves)
  - Cancer immunotherapies (due to its MOA)
- Diminishing effect (merging curves)
  - Overall survival confounded by subsequent therapies

Delayed Effect (OAK)



Rittmeyer, A., et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial (2017) The Lancet, 389 (10066), pp. 255-265.

Diminishing Effect (BOLERO II)



Piccart M., et al. Everolimus plus exemestane for hormone-receptor-positive, human epidermal growth factor receptor-2-negative advanced breast cancer: overall survival results from BOLERO-2. Ann Oncol. 2014 Dec;25(12):2357-6

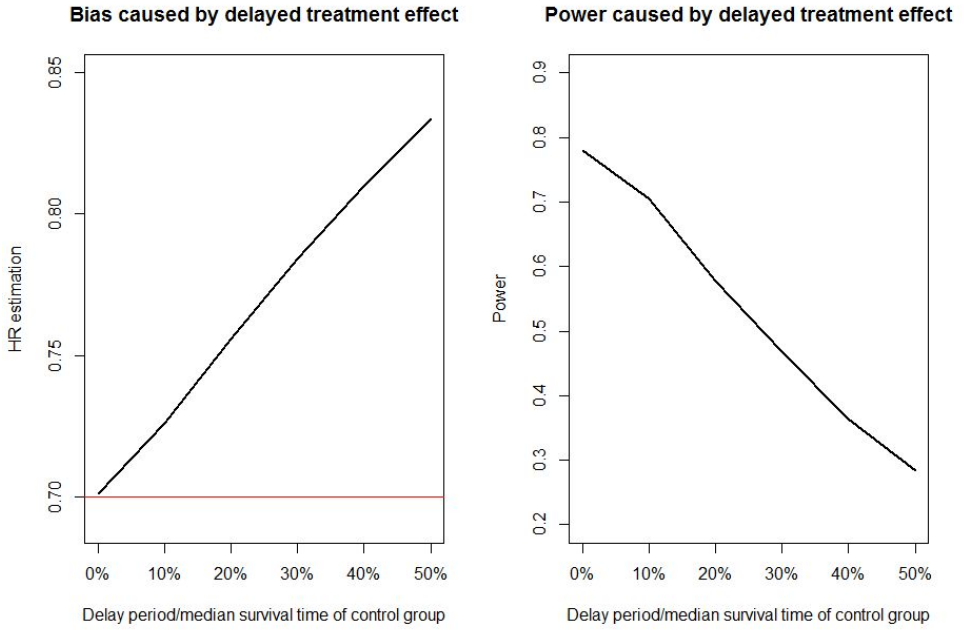
Impact on Analysis

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

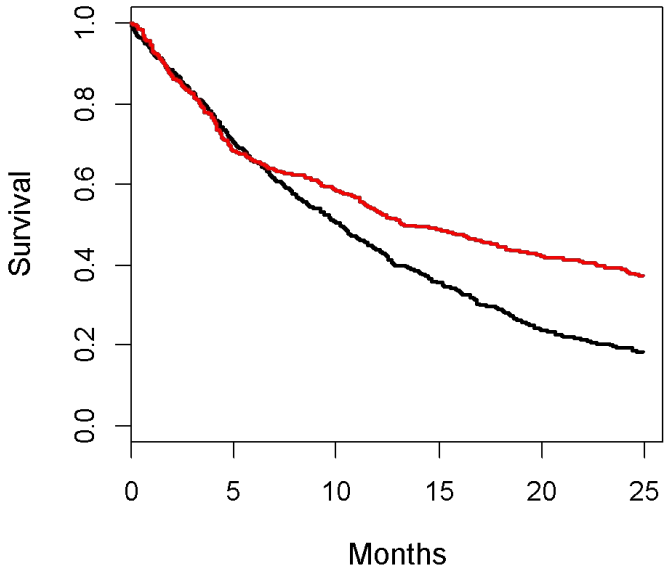
Delayed treatment effect

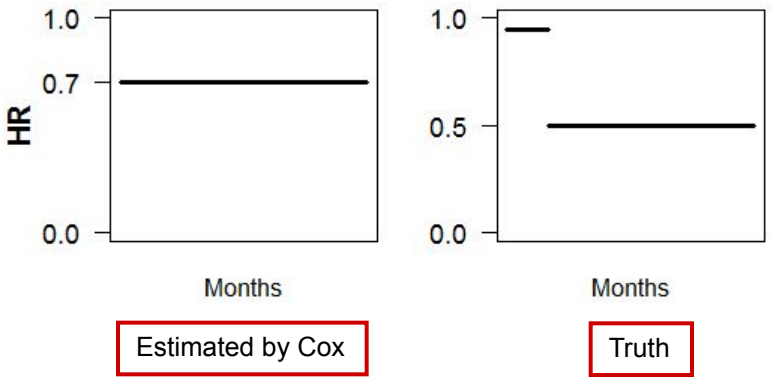
$$\text{HR}=\begin{cases}1 & for\ t \le t^0\\ 0.7 & for\ t > t^0\end{cases}$$



Impact on Clinical Interpretation

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 can not stay on treatment for a sufficient time may be unlikely to get benefits
2. Evaluation for health economics



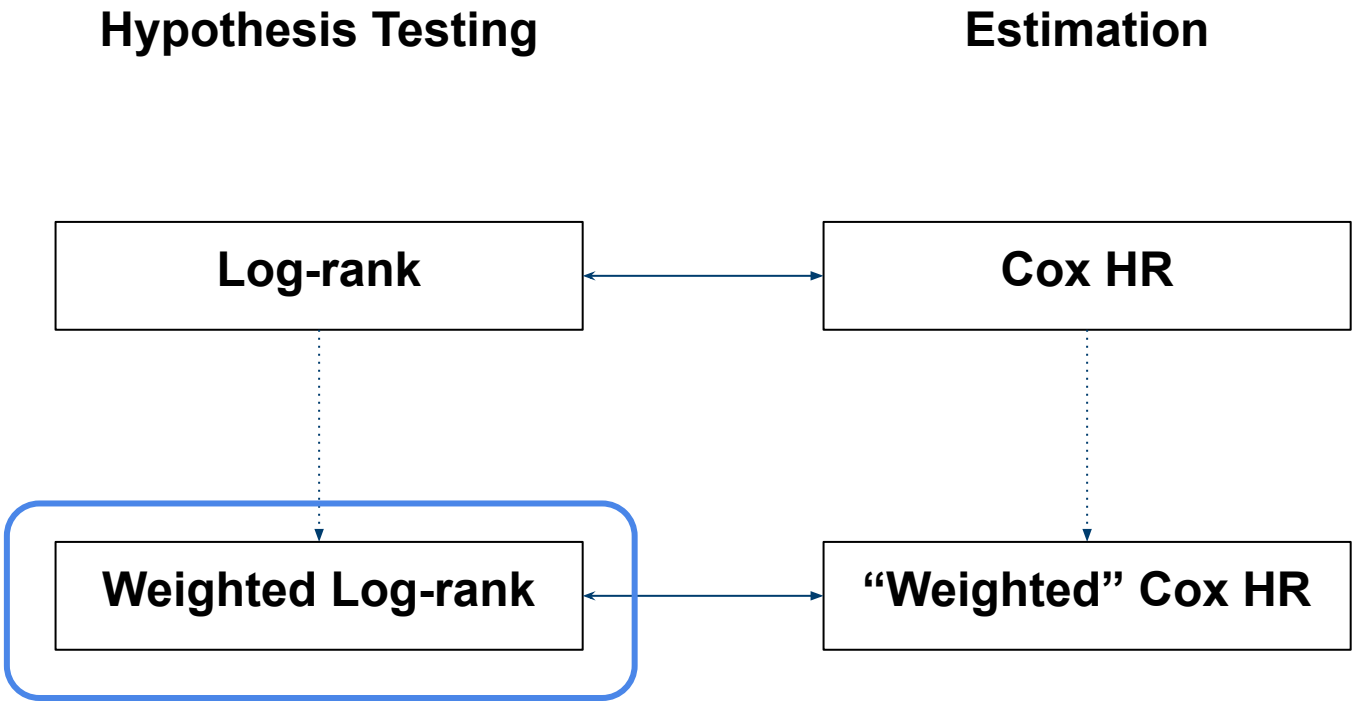


Efforts on NPH

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
- Milestones:
  - Duke Margolis workshop Feb. 5, 2018 with FDA and EMA participants
  - Two manuscripts under review
    - Alternative analysis methods for time to event endpoints under non-proportional hazards: a comparative analysis
    - Robust design and analysis of clinical with trial non-proportional hazards: a straw man guidance
- Members
  - AZ, BMS, Merck, B&I, Novartis, Eli Lilly, Abbvie, Roche/Genentech, Bayer, Janssen, Takeda, Amgen, Pfizer, GSK, Celgene, Sanofi

Overview of Analysis Methods

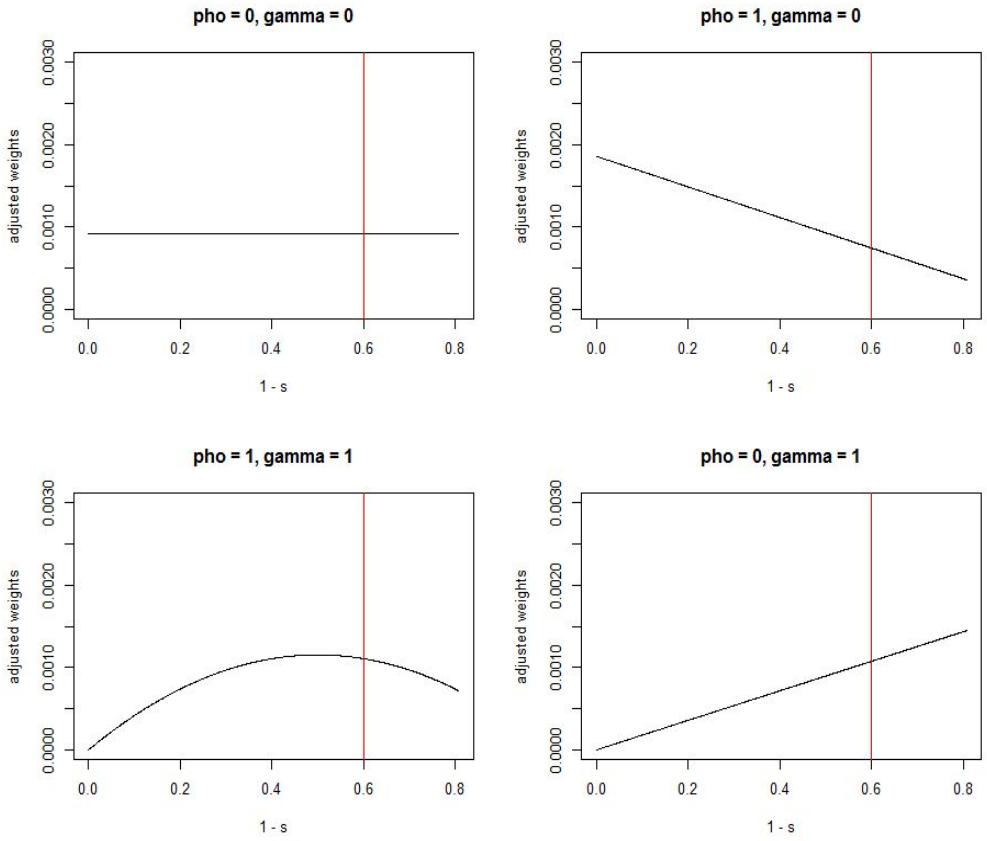


Weighted Log-Rank Tests

- 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

Fleming-Harrington weight family



MaxCombo Test

- Pre-specify a set of K weight functions and select the one with the best Z value
- P-value adjusted to preserve Type-I error
- Examples
  - WG[1]: FH(0,0), FH(1,0), FH(1,1), FH(0,1)
  - Lee[2]: FH(1,0), FH(0,1)

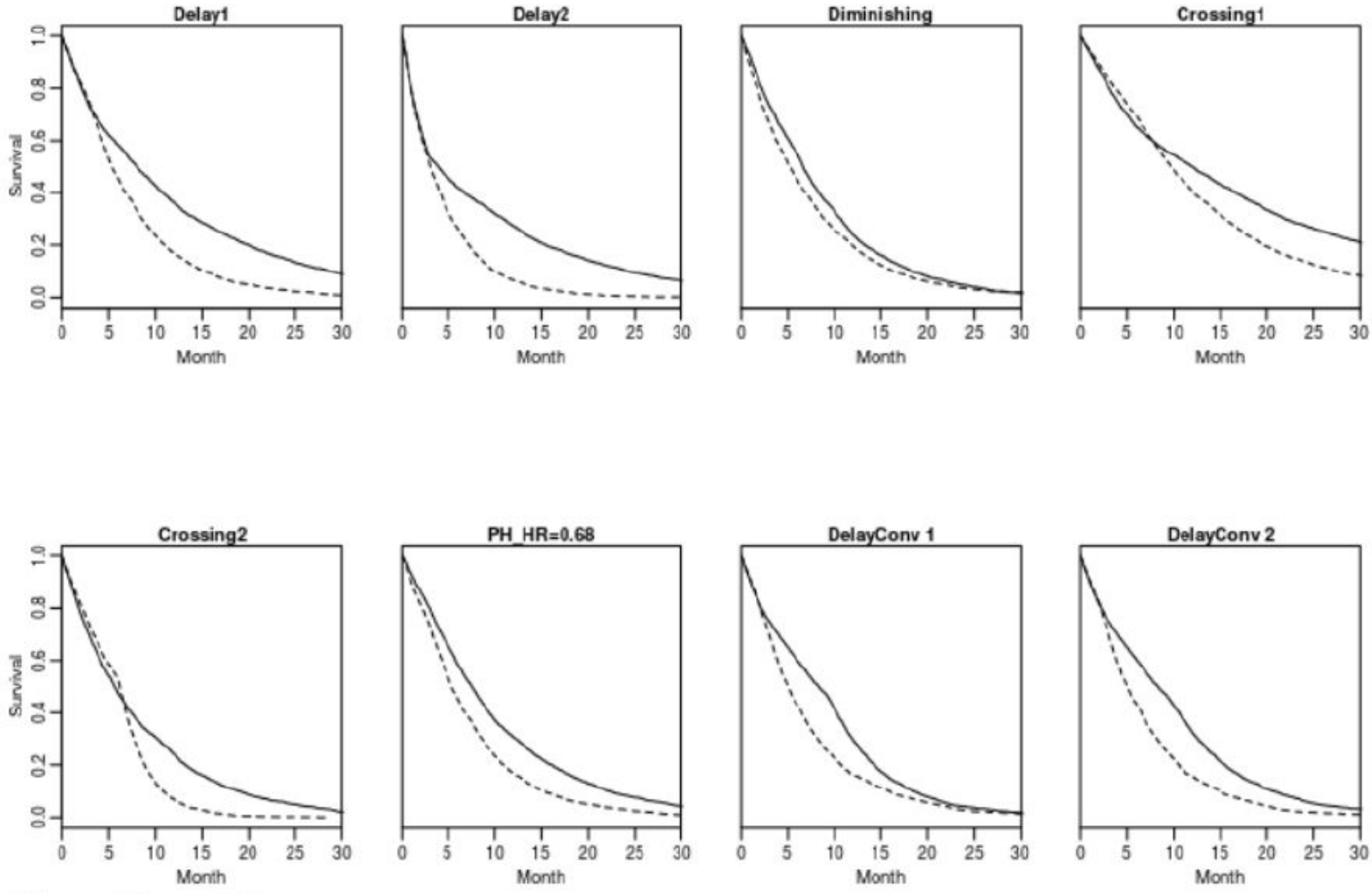
1. Lin, R.S. et.al Alternative analysis methods for time to event endpoints under non-proportional hazards: a comparative analysis. Under review.
2. S Roychoudhury. et al. Robust design and analysis of clinical with trial non-proportional hazards: a straw man guidance. Under review.
3. Lee, S-H. (2007). On the versatility of the combination of the weighted log-rank statistics. Computational Statistics & Data Analysis 51:6557-6564.

Simulation Study #1 (NPH WG)

- Settings
  - 9 patterns (8 NPH/PH, null)
    - delayed effect, diminishing effect, both
  - 3 Categories of tests
    - Rank-based tests
      - Log-rank, weighted log-rank
    - Kaplan-Meier curve-based tests
      - Weighted KM, RMST
    - Combination tests
      - Breslow test, MaxCombo test

Reference: Lin, R.S. et.al Alternative analysis methods for time to event endpoints under non-proportional hazards: a comparative analysis. Under review.

NPH/PH Patterns



Simulation Study #1 Summary

- Testing
  - Type I error is preserved
  - No test achieves the highest power across all scenarios
  - MaxCombo has relatively robust power across patterns

- HR for MaxCombo (Sasieni 1993)
  - Less biased than Cox HR
  - Slightly biased (anti-conservative) under null and PH due to model selection

Reference: Lin, R.S. et.al Alternative analysis methods for time to event endpoints under non-proportional hazards: a comparative analysis. Under review.

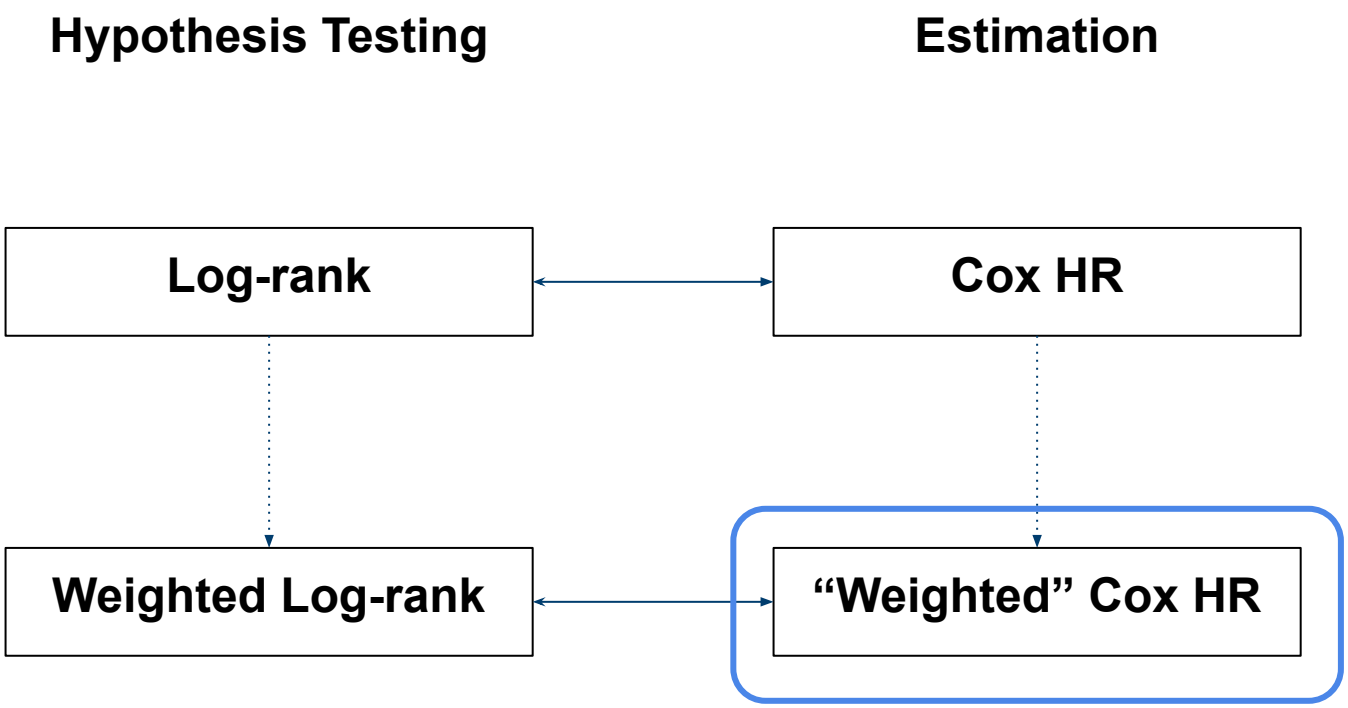
Clinical Interpretation?

- What is the clinical interpretation of the weights?

- Effect size changes over time under NPH
  - “Time-profile”
  - Cannot be summarized by a single measure

We proposed a hazard ratio time-profile estimated via “weighted” Cox

Overview of Analysis Methods



1. Clinical interpretation of the weights
2. HR "time-profile"

"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-varying covariate)
    - Create Y, fit the Cox model and get the coefficient estimate for Y (ie, β)
  - Time-Varying HR (time-profile)
    - $HR(t)=\exp[\beta*W(t)] = [HR_F]^{W(t)}$

Full effect HR_F

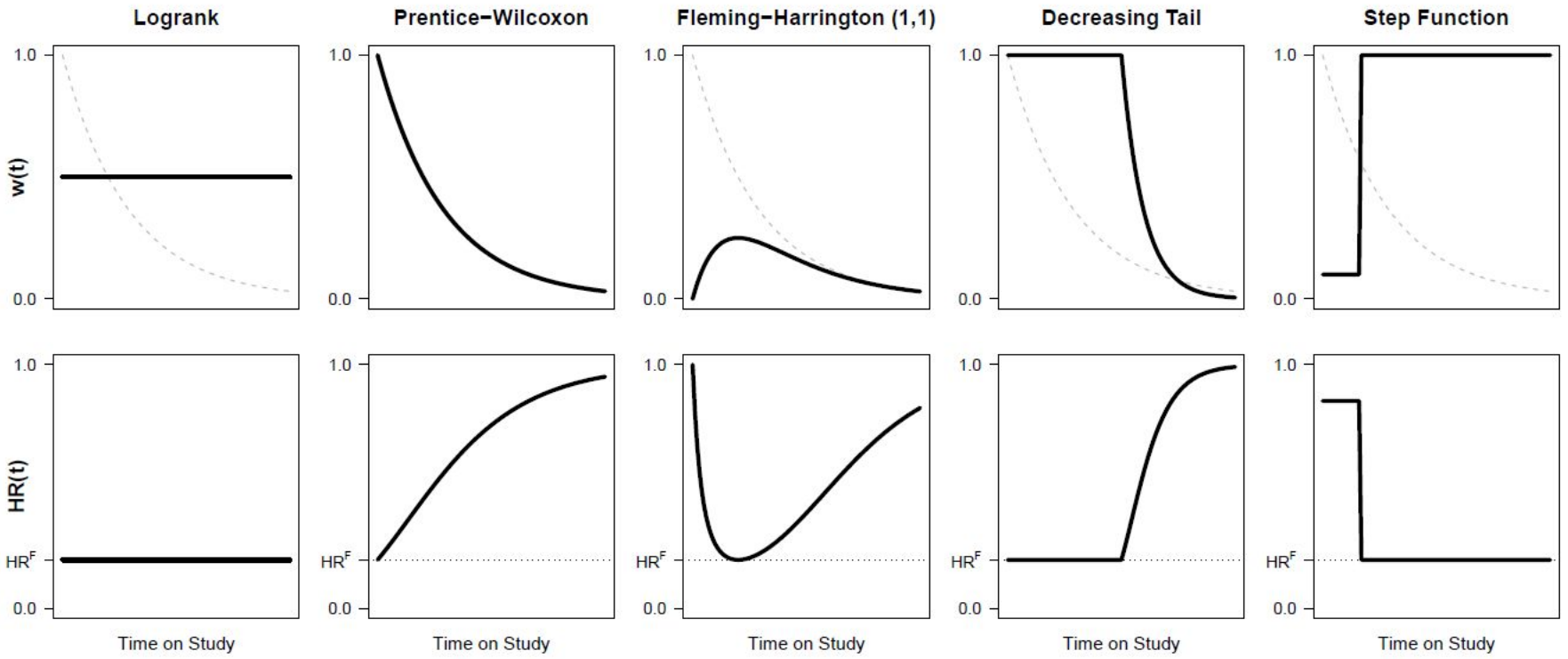
Effect adjusted by W(t)

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

Reference: Lin, R.S. and Leon L.F., Estimation of treatment effects in weighted log-rank tests. Contemporary Clinical Trials Communications, Volume 8, 2017, Pages 147-155

Examples of Weight Functions & HR Time-Profiles

- The shape of the HR time-profile is determined by the weight function
- The full effect HR_F is estimated from the data

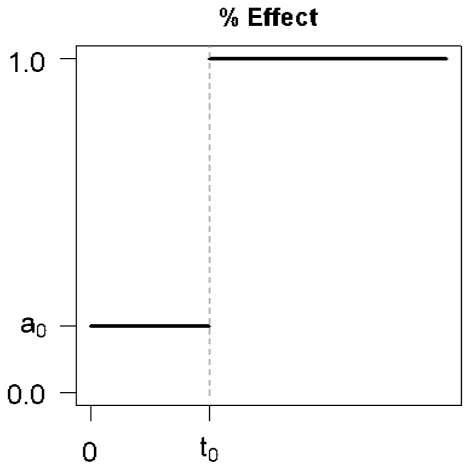


Choice of weight functions

- Predictable
  - Asymptotic properties derived from the martingale framework
- Optimal?
  - $w(t) \propto \log(HR(t))$
  - unknown and hard to estimate
- Meaningful clinical interpretation
  - FH(1,1) or FH(1,0)
    - Full effect takes place at only one time point
      - Not useful for clinical interpretation
      - Over-estimate the actual effect
  - A long period with w(t)=1
    - E.g., piecewise constant weights

Simulation Study #2
Delay Effect

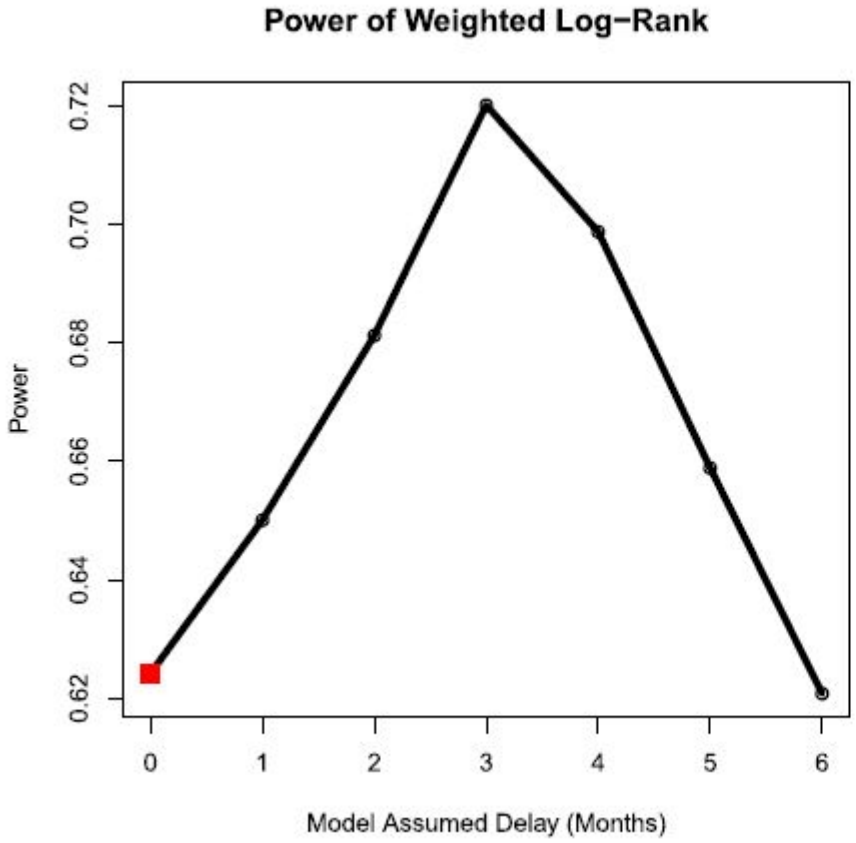
- Settings
  - N = 400 (1:1 randomization)
  - Event-patient ratio 70%
  - Treatment effect
    - HR=0.68 after 3 months
    - HR=0.9 (27%) during 0-3 months
  - Control median survival: 12 mos
  - Varying
    - Delay duration (t_0)
    - Adjustment factor (a_0)
  - Enrollment : 12 mos with ramp-up
  - 10,000 runs

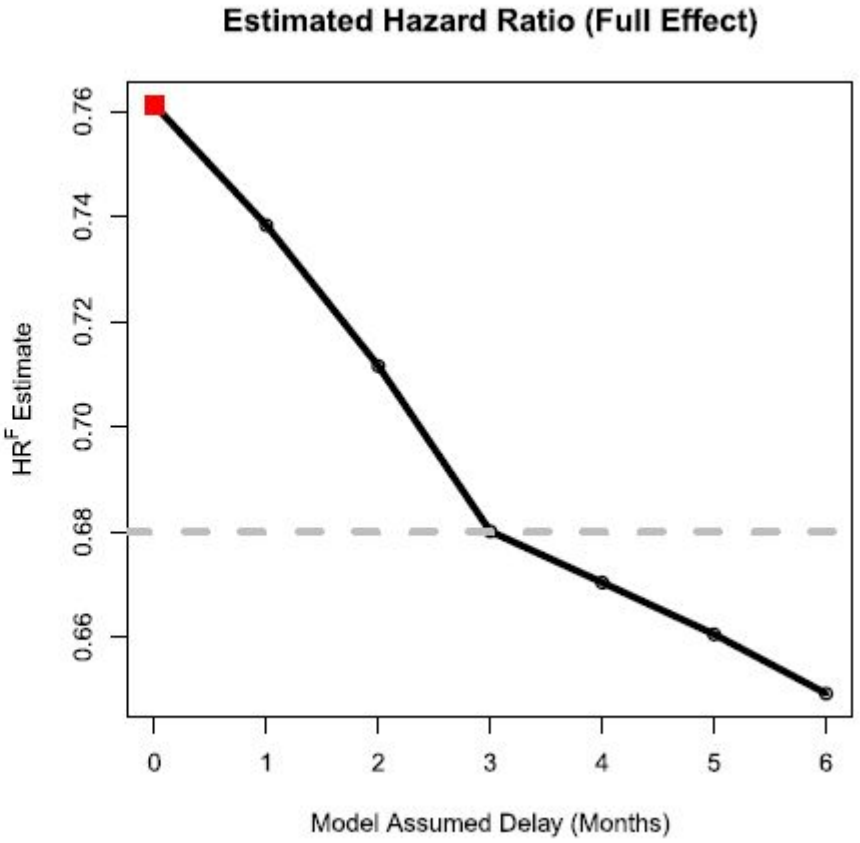


Reference: Lin, R.S. and Leon L.F., Estimation of treatment effects in weighted log-rank tests. Contemporary Clinical Trials Communications, Volume 8, 2017, Pages 147-155

Power and HR Estimates Varying t_0

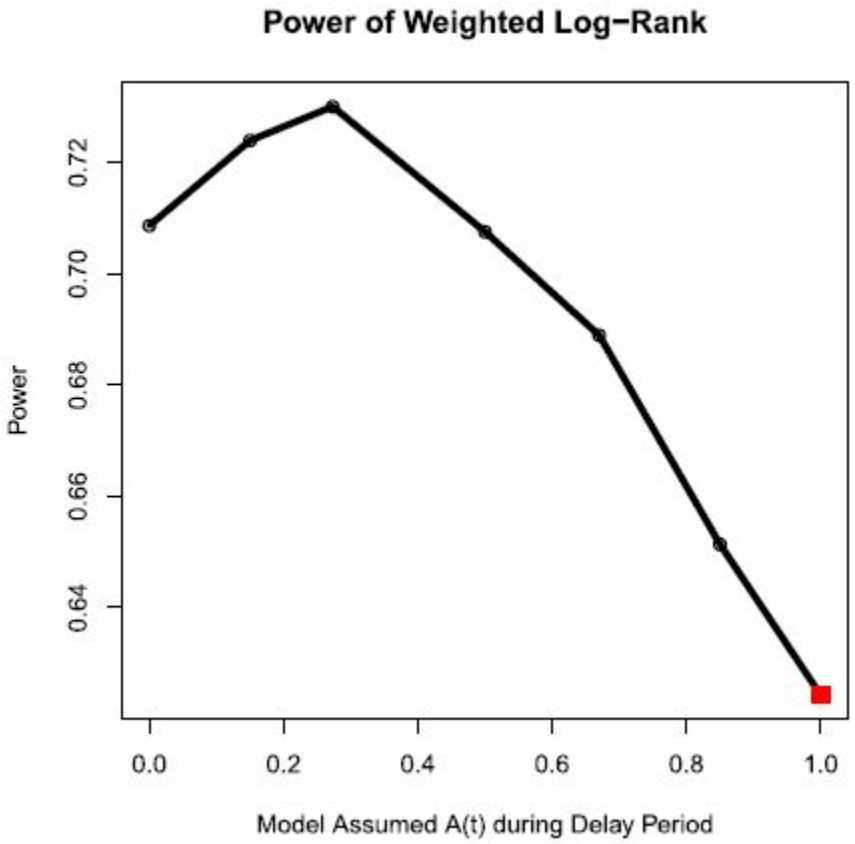
True t_0 = 3 months; a_0 = 27%

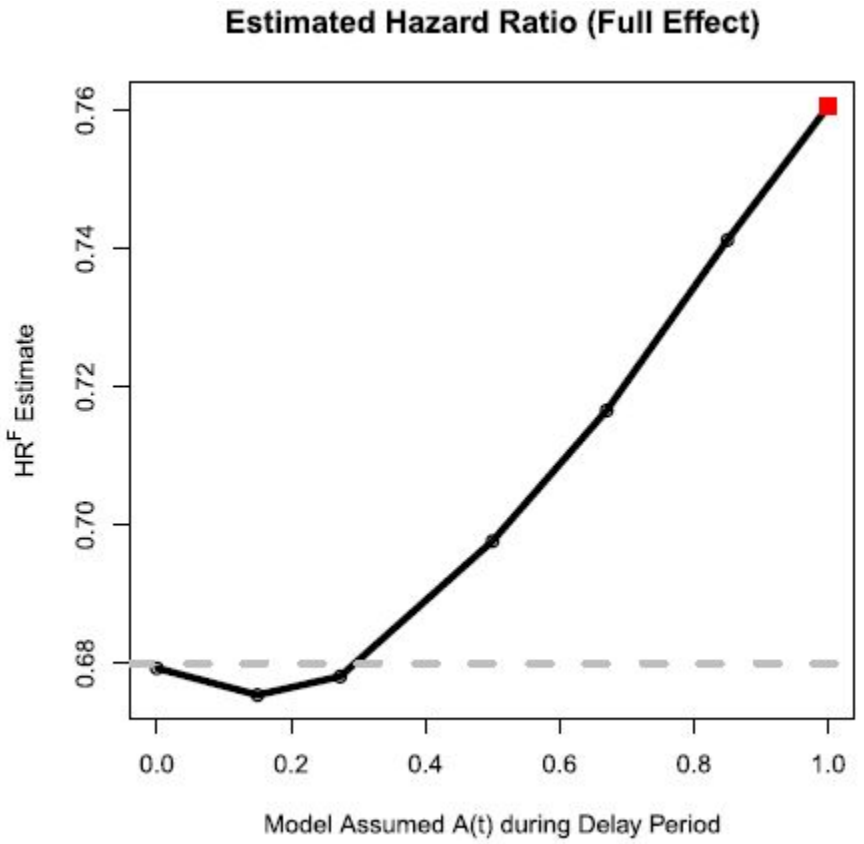




Power and HR Estimates Varying a_0

True t_0 = 3 months; a_0 = 27%



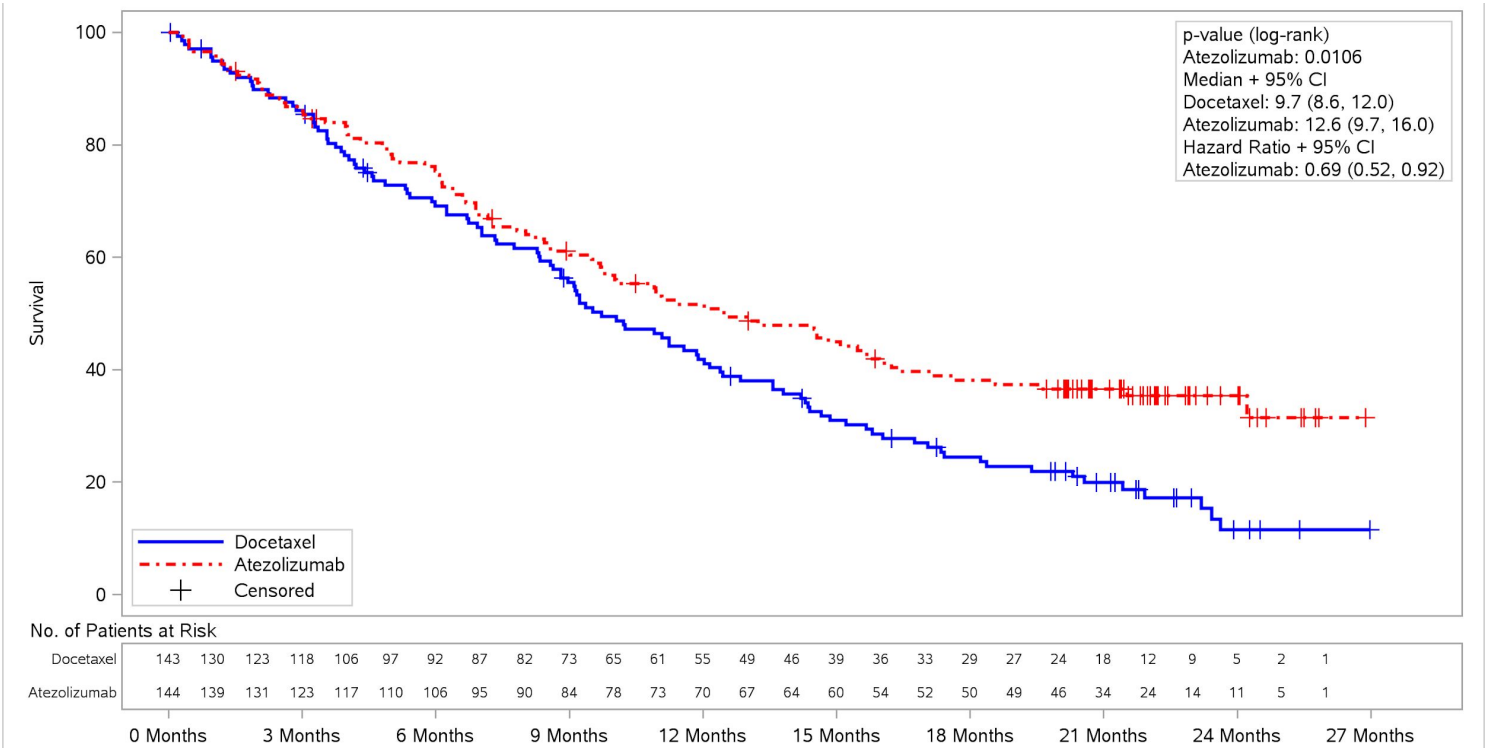


Simulation Study #2 Summary

- Type-I error is preserved
- Power
  - The most powerful test if the weight assumption is correct
  - In general more powerful than standard log-rank tests even if mis-specified
- Hazard ratio estimate
  - Unbiased if the assumption is correct
  - In general less biased than standard Cox models even if mis-specified

Reference: Lin, R.S. and Leon L.F., Estimation of treatment effects in weighted log-rank tests. Contemporary Clinical Trials Communications, Volume 8, 2017, Pages 147-155

Example: Poplar 2L NSCLC



Effect Time-profile	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)

•*Unstratefied analysis based on data cut Dec 1, 2015

Summary

- Under NPH, “weighted” Cox HRs could be reported in addition to standard Cox HR
  - Connected to the weighted 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

- Main collaborator
  - Larry Leon
- NPH Cross-Pharma Working Group
  - Leadership team
    - Tai-Tsang Chen
    - Renee Iacona
    - Ji Lin
    - Keaven Anderson
    - Pralay Mukhopadhyay
    - Satrajit Roychoudhury
  - Members
    - Bo Huang, Jason Liao, Rong Liu, Xiaodong Luo, Rui Qin, Kay Tatsuoka, Xuejing Wang, Yang Wang, Jian Zhu, Honglu Liu, Tianle Hu, Prabhu Bhagavatheeswaran. Julie Cong, Margarida Geraldes, Dominik Heinzmann, Yifan Huang, Zhengrong Li, Honglu Liu, Yabing Mai, Jane Qian, Li-an Xu, Jiabu Ye, Luping Zhao
- Genentech/Roche
  - Jing Yi, Ru-Fang Yeh, Ben Lyons, Gracie Lieberman, Yumeng Li, Shi Li, Na Cui, Ina Rhee, Cancer Immunotherapy Endpoint Taskforce

Q&A