Simulations for Non-Proportional Hazards

Acknowledgement

- Team Leads:
  - Tianle Hu (Eli Lilly)
  - Satrajit Roychoudhury (Pfizer)
  - Keaven Anderson (Merck)
  - Julie Cong (Boehringer Ingelheim)

- Special Thanks to
  - Larry Leon (Genentech)
  - Honglu Liu (Eli Lilly)

- Team includes members from:
  - Abbvie
  - AstraZeneca
  - Bayer
  - BMS
  - Boehringer Ingelheim
  - Genentech/Roche
  - Johnson & Johnson
  - Merck
  - Pfizer
  - Sanofi
  - Takeda

Simulation Scope

Methods

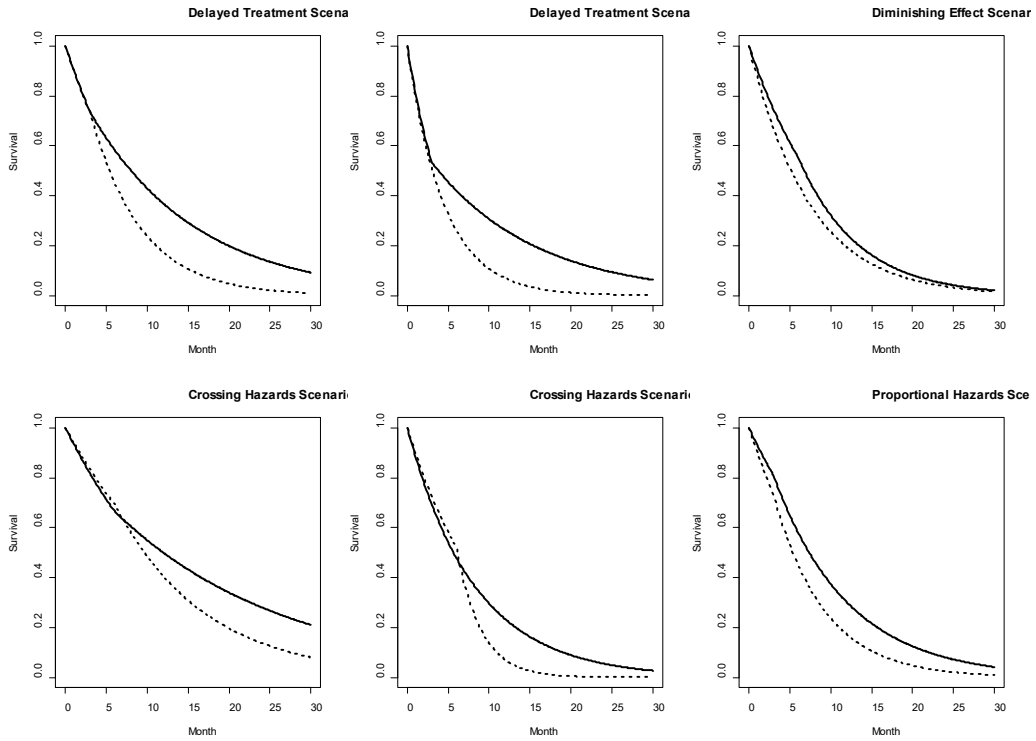
- Category 1: LR, FH(0,1), FH(1,0), FH(1,1)
- Category 2: Weighted K-M, RMST
- Category 3: Breslow combo, max combo

Scenarios

- Delayed Treatment 1 & 2
- Diminishing Effect (Belly Shape)
- Crossing Hazards 1 & 2
- Proportional Hazards
- Null Scenario

- ❖ Delay scenarios are informed by CM – 141 (2L SCCHN) and CM – 017 (2L Squamous NSCLC)
- ❖ Diminishing effect scenario is is informed by AVAGAST Trial (1L Gastric Cancer)
- ❖ Crossing Hazards are informed by CM- 057 (2L Non-squamous NSCLC) and IPASS (1L NSCLC)

Simulation Setup



Sample size: 300, 600 and 1200.
Enrollment Duration: 12 mos, 18 mos and 24 mos.
Dropout hazard rate: λ=0.014.
Number of events: 210

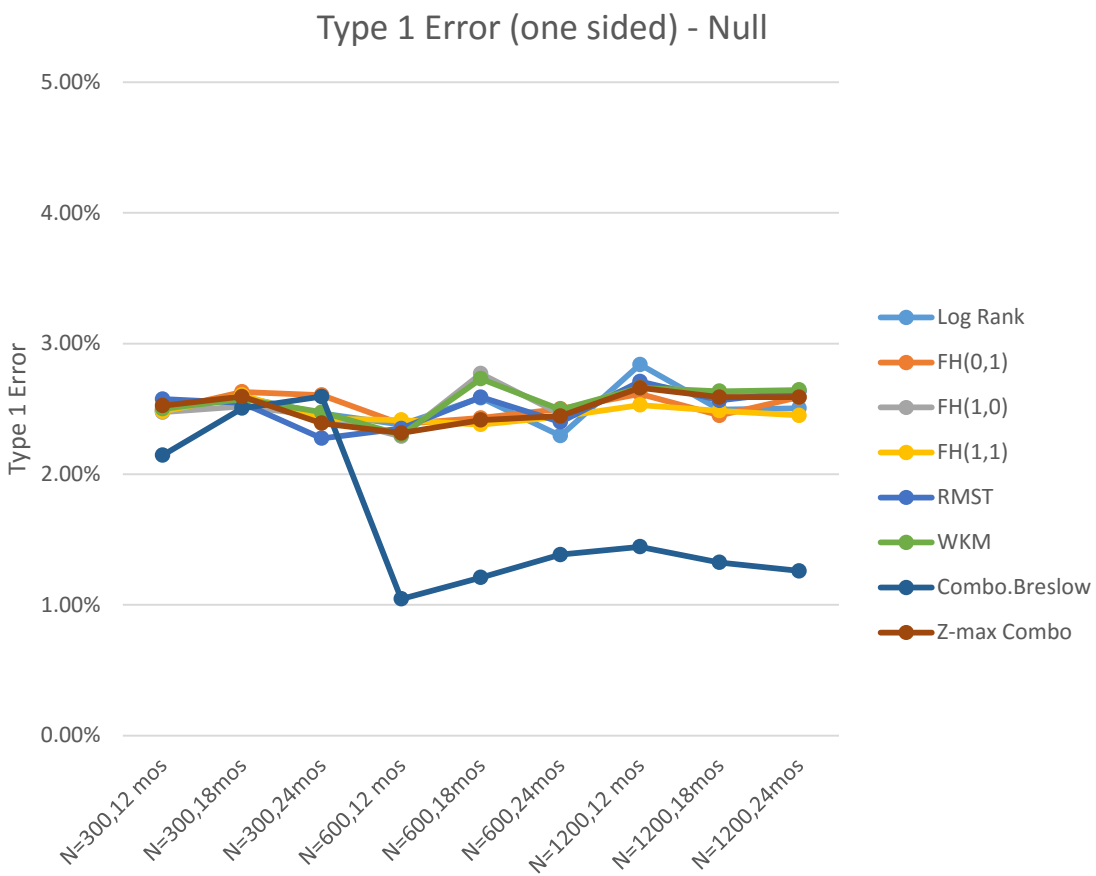
Scenario	CP	0 ≤ t < CP			t ≥ CP		
		λC1	λE1	HR1	λC2	λE2	HR2
Delayed Effect 1	3	0.104	0.103	0.990	0.161	0.077	0.478
Delayed Effect 2	3	0.226	0.210	0.929	0.222	0.079	0.356
Diminishing Effect	6	0.134	0.098	0.731	0.140	0.137	0.979
Crossing Hazards 1	6	0.061	0.068	1.115	0.090	0.048	0.533
Crossing Hazards 2	6	0.108	0.123	1.139	0.334	0.120	0.359
Proportional Hazards	3	0.104	0.071	0.680	0.161	0.110	0.680
Null	3	0.104	0.104	1.000	0.161	0.161	1.000

Cases	events =70%*300	events =35%*600	events =17.5%*1200
12 months	N=300,12mos	N=600,12mos	N=1200,12mos
18 months	N=300,18mos	N=600,18mos	N=1200,18mos
24 months	N=300,24mos	N=600,24mos	N=1200,24mos

Testing

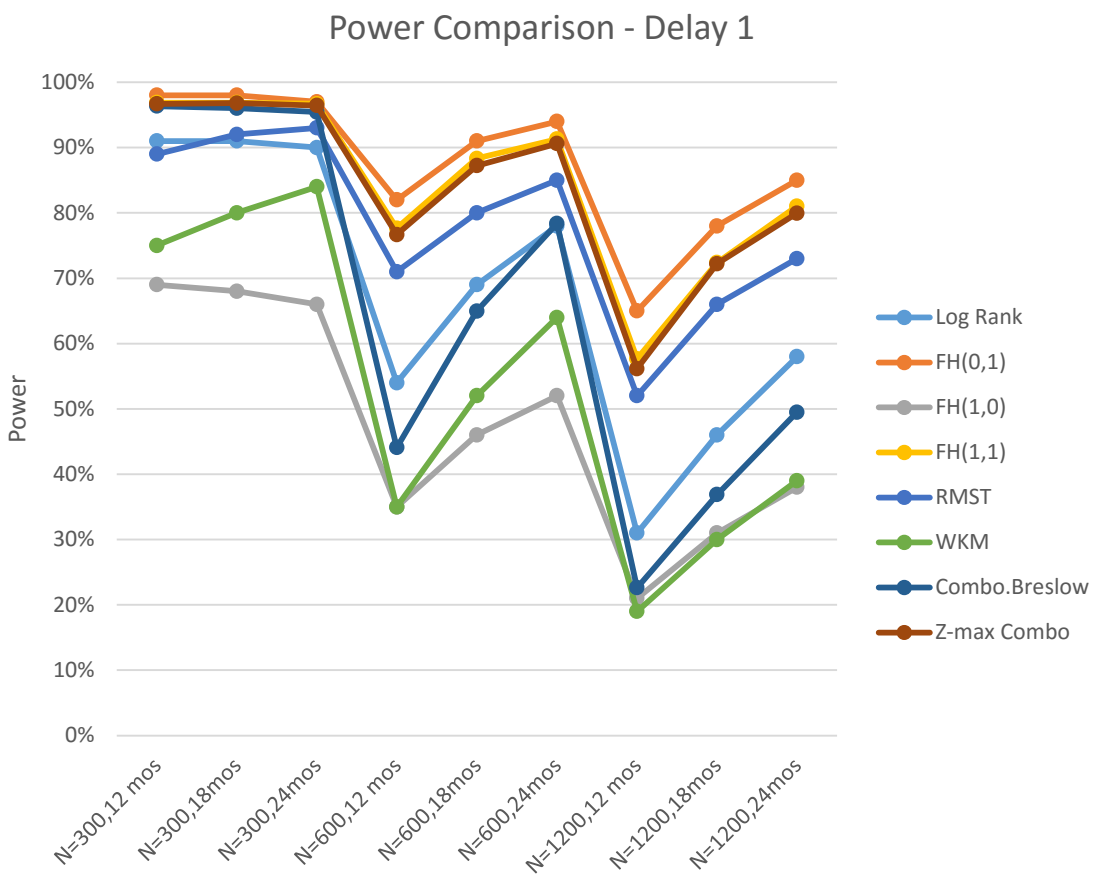
Null Scenario (Type 1 Error)

- All methods control type 1 error well across cases.
- There are random spikes over 2.5%, but mostly within simulation standard error (0.1% based on 20,000 iterations).
- Combo.Beslow requires asymptotic independence:
  - In finite sample, independence assumption may not hold.



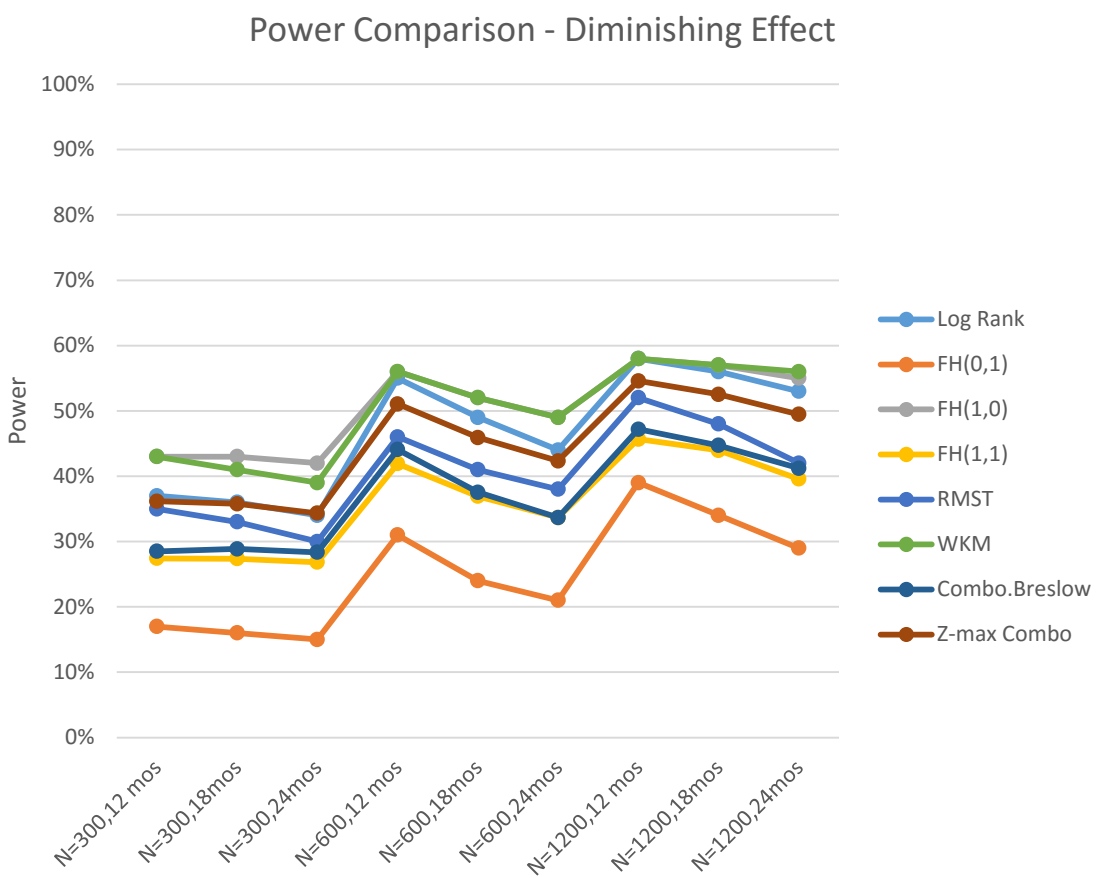
Delayed Scenario

- The max combo test has clear advantage over LRT in terms of power.
- In fact, its performance is close to that of FH(0,1), which is expected to perform well.
- The advantage is larger with higher censoring.
- The K-M based test statistics don't perform as well.
- Irrespective of tests, the power increases with maturity and enrollment time.



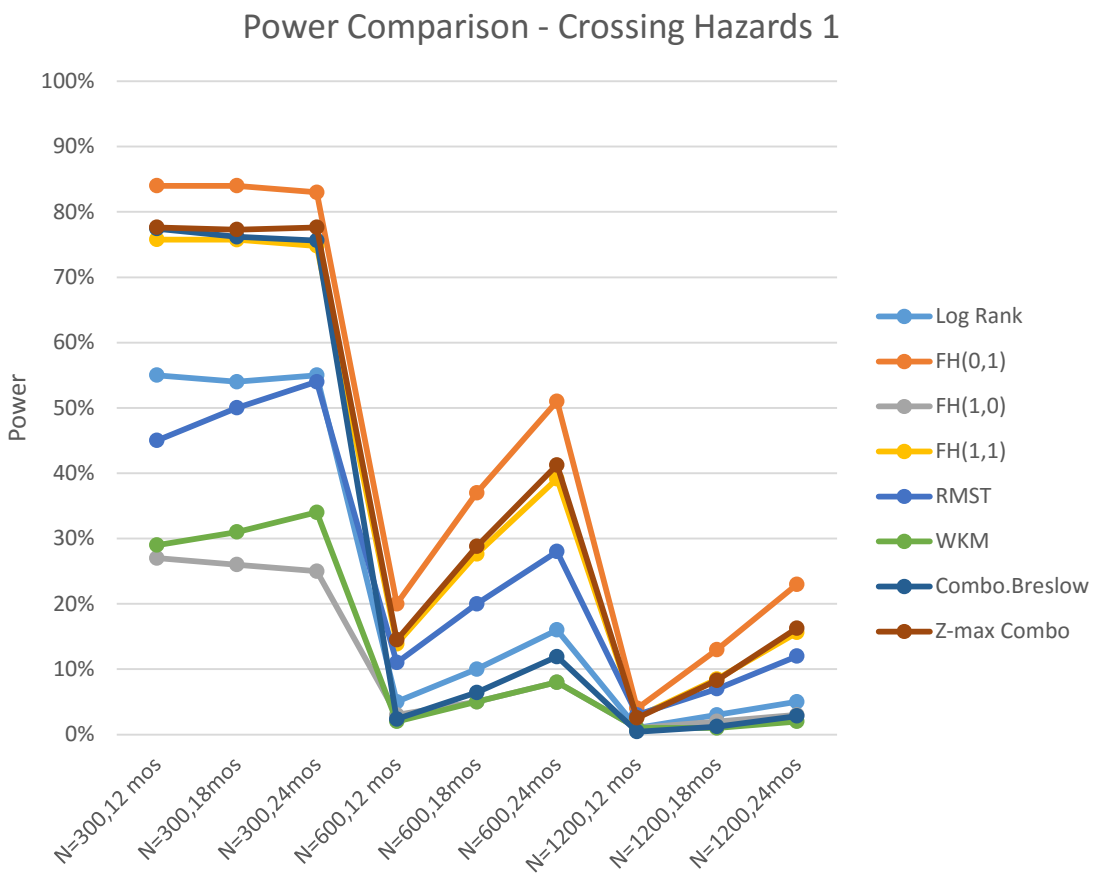
Diminishing Scenario

- Diminishing is a challenging scenario because the “overall” treatment effect is usually small
- Under the diminishing effect, max combo has ~4% less power than LRT.
- The FH(1,0) does better, but not by a whole lot.
- Weighted K-M does well too, similar to FH(1,0)
- Irrespective of tests, the power decreases with enrollment time and maturity.



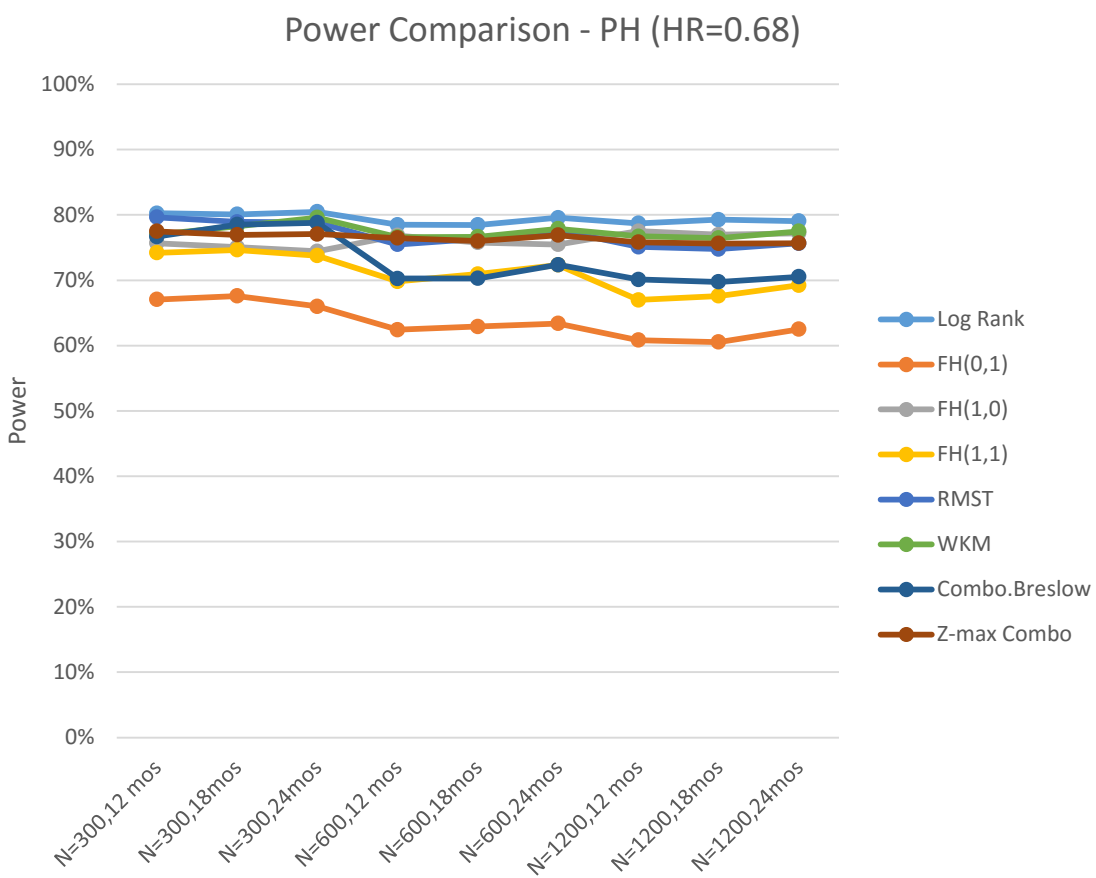
Crossing Hazards Scenario

- Crossing hazards ($HR_1>1$ and $HR_2<1$) is very similar to delayed effect
- The max combo test has a clear advantage over LRT in terms of power.
- In fact, its performance is close to that of FH(0,1), which is expected to perform well.
- The K-M based test statistics don’t perform as well.

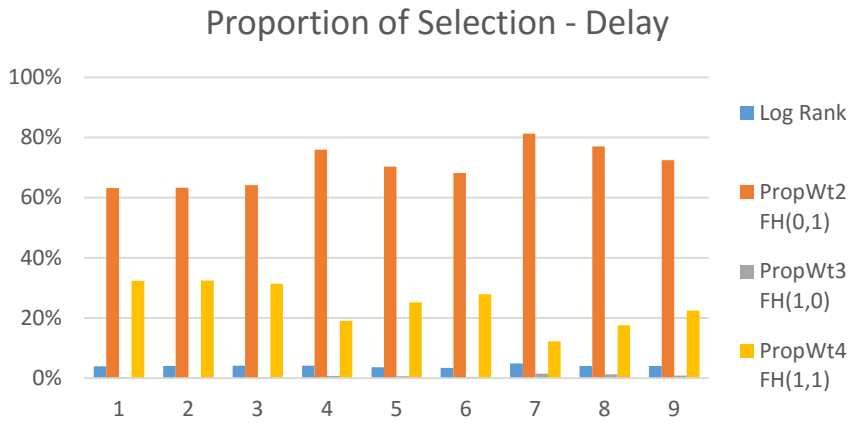


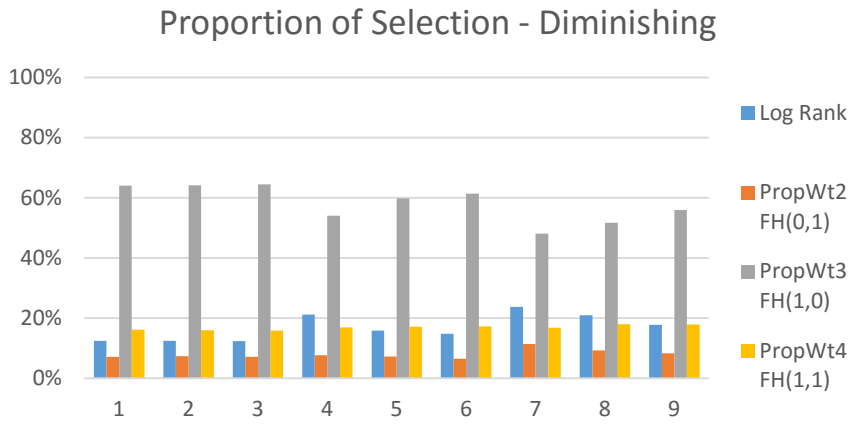
PH Scenario

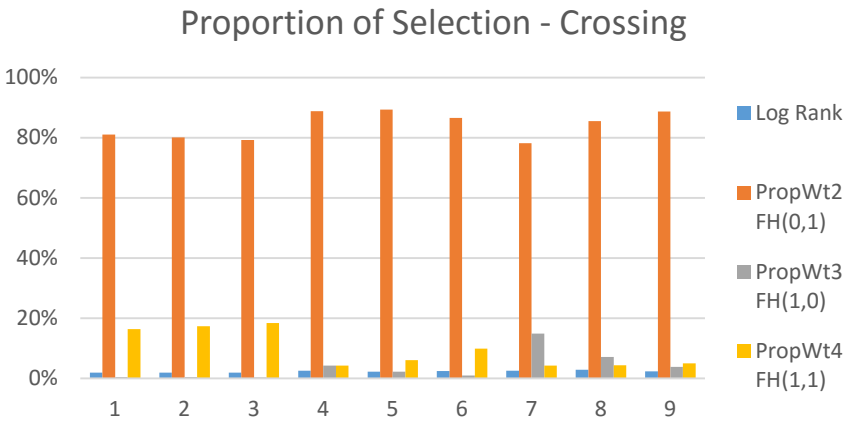
- The LRT is the semi-parametric most efficient under PH.
- However, most of the tests we considered are quite competitive:
  - mostly within 10% power difference, except FH(0,1).
- max combo is about 3-4% power inferior to the LRT.
- Power only depends on the number of events

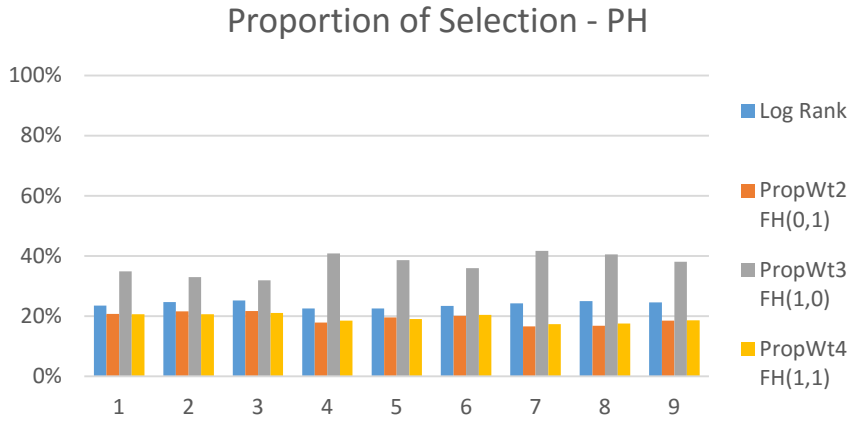


Model Selection Probabilities of Max-combo



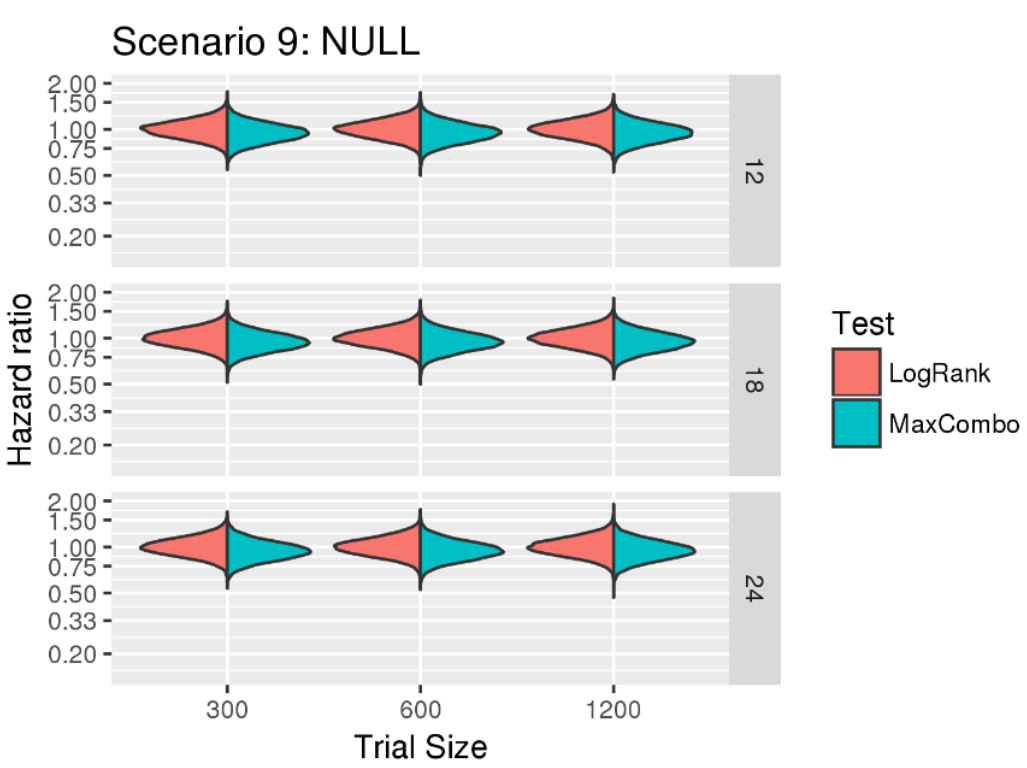


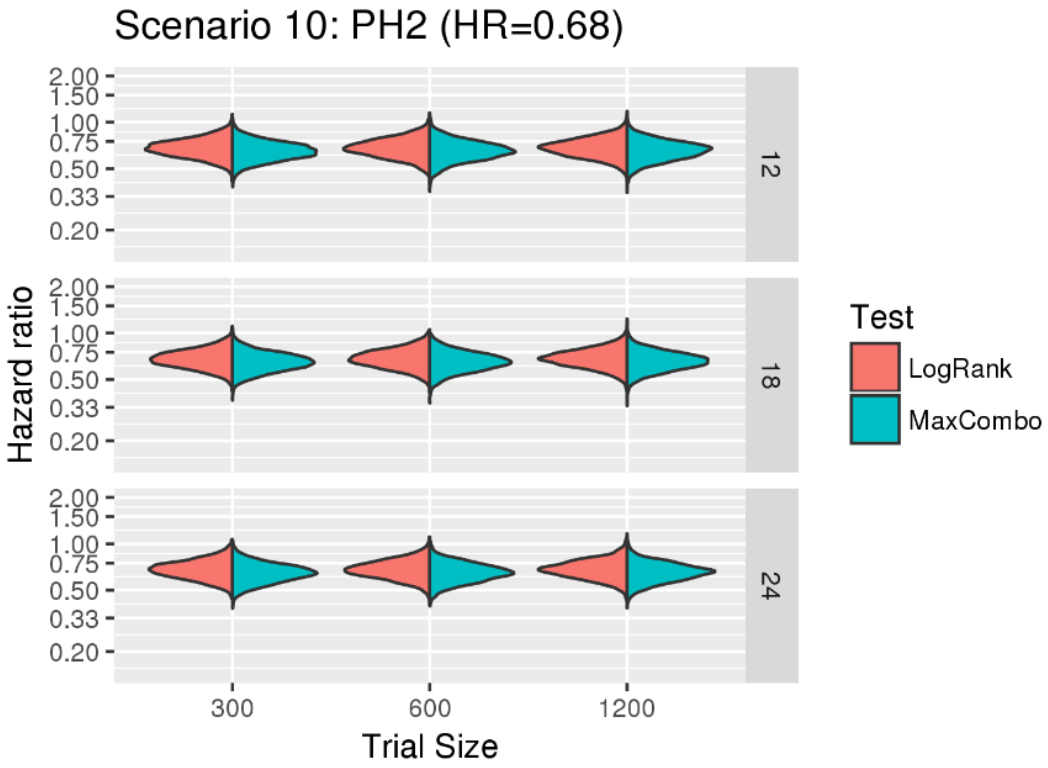




Estimation

HR vs. Max-combo Estimate (Null and PH)

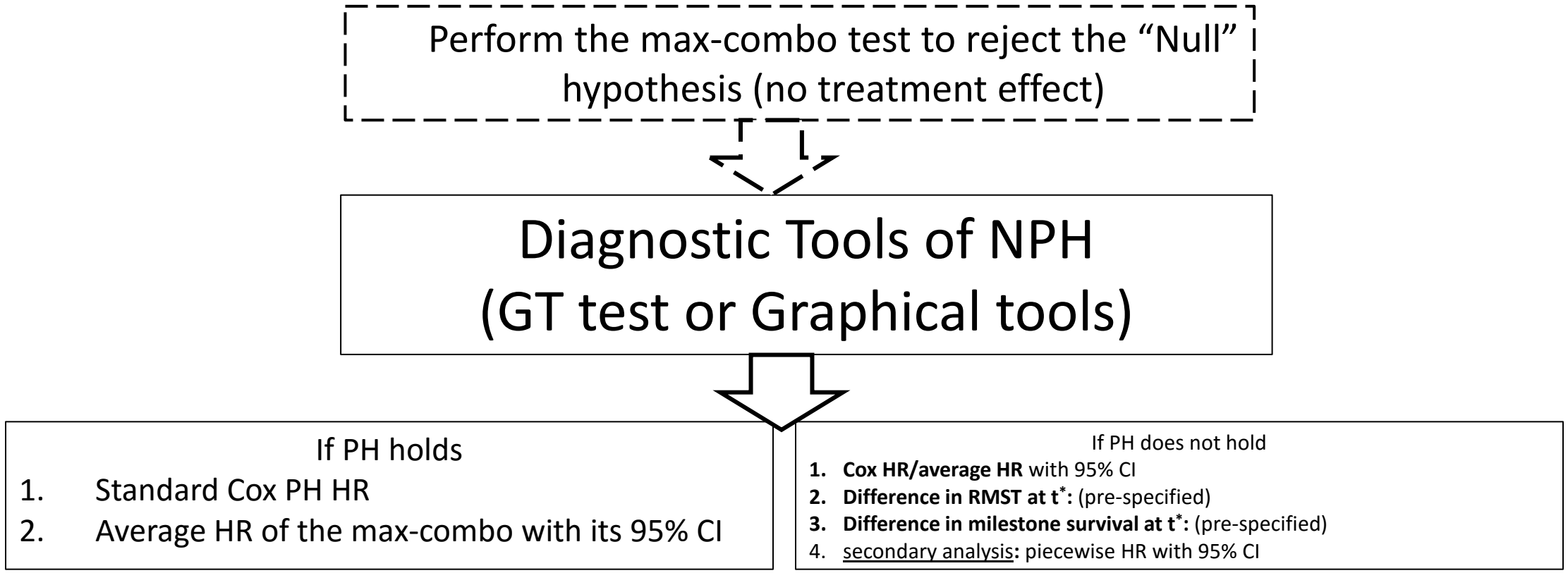




Max-combo Estimate

- Based on the simulation, with the limited set of tests in the max combo test, the bias in the point estimate is negligible.
  - under the null (HR=0.95 vs true HR of 1)
  - under the PH (HR=0.65 vs true HR of 0.68)
- In all other scenarios, the point estimate of the max.combo is the average HR, which depends on the weight function.

Procedure to Estimate Trt Effect



Will illustrate this procedure in Session III

Conclusions and Recommendations

- Max combo test is robust and agnostic to the types of non-PH:
  - A very strong upside under delayed effect or crossing hazards scenarios (both quite commonly being observed within IO)
  - Acceptable loss in power under PH and diminishing effect (3-4%).
  - Such trade-off motivates the max combo to be a competitive test.
- Effect estimation under NPH is complex.
  - Max-combo estimate: the bias is negligible under the null and PH.
  - For treatment effect estimate, take a data-dependent approach
    - If PH according to the diagnostic tool (GT test or graphical tool), then report regular HR and max-combo estimate.
    - Otherwise, RMST difference, milestone rates and piecewise HR in addition.

Questions to the Panel

1. Do you agree max-combo is an appropriate **test** when the trialist is uncertain of NPH?
2. Do you agree max-combo **estimate** is an useful measure of treatment effect?
3. Is a data dependent approach to estimating the treatment effect acceptable?