CONSIDERATIONS FOR DESIGN AND ANALYSIS OF TRIALS WITH POSSIBLY NON-PROPORTIONAL HAZARDS

October 16, 2018

Keaven M. Anderson, Ph.D. and Satrajit Roychoudhury

Merck Research Laboratories and Pfizer

Disclaimer

While the authors are members of the Non-Proportional Hazards (NPH) Working Group, any mistakes and opinions should be considered those of the authors. Also, this work does not represent a company position for either Merck or Pfizer.

Acknowledgements

- **Members of the Cross Pharma NPH Working Group**

  - **Leadership Team**
    - **CoLeaders:** Renee Iacona (AZ), Tai-Tsang Chen (BMS)
    - **Design and Analysis Workstream:** Keaven Anderson (Merck), Satrajit Roychoudhury (Pfizer), Tianle Hu/ (Lilly), Ray Lin (Roche)
    - **Endpoint Workstream:** Jane Qian (Abbvie), Dominik Heinzmann (Roche)
    - **Simulation Team:** Julie Cong (B&I), Tianle Hu (Lilly)
    - **Case Studies:** Pralay Mukhopadhyay (AZ)

  - **Organizations represented in the Working Group**
    - **AZ, BMS, Merck, Boehringer Ingelheim, Novartis, Lilly, Abbvie, Genentech, Roche, Bayer, Janssen, Takeda, Amgen, Pfizer, GSK, Celgene, Sanofi, Johnson & Johnson, and FDA**

Outline

- Background
- Methods studied
- Simulation summary
- Recommendations for testing and estimation
- Design considerations
- When do the results break down?
- Summary

MERCK

BACKGROUND

Based on slides from Rajeshwari Shridhara, FDA

https://healthpolicy.duke.edu/events/public-workshop-oncology-clinical-trials-presence-non-proportional-hazards

Time to Event Analysis in Randomized Clinical Trials

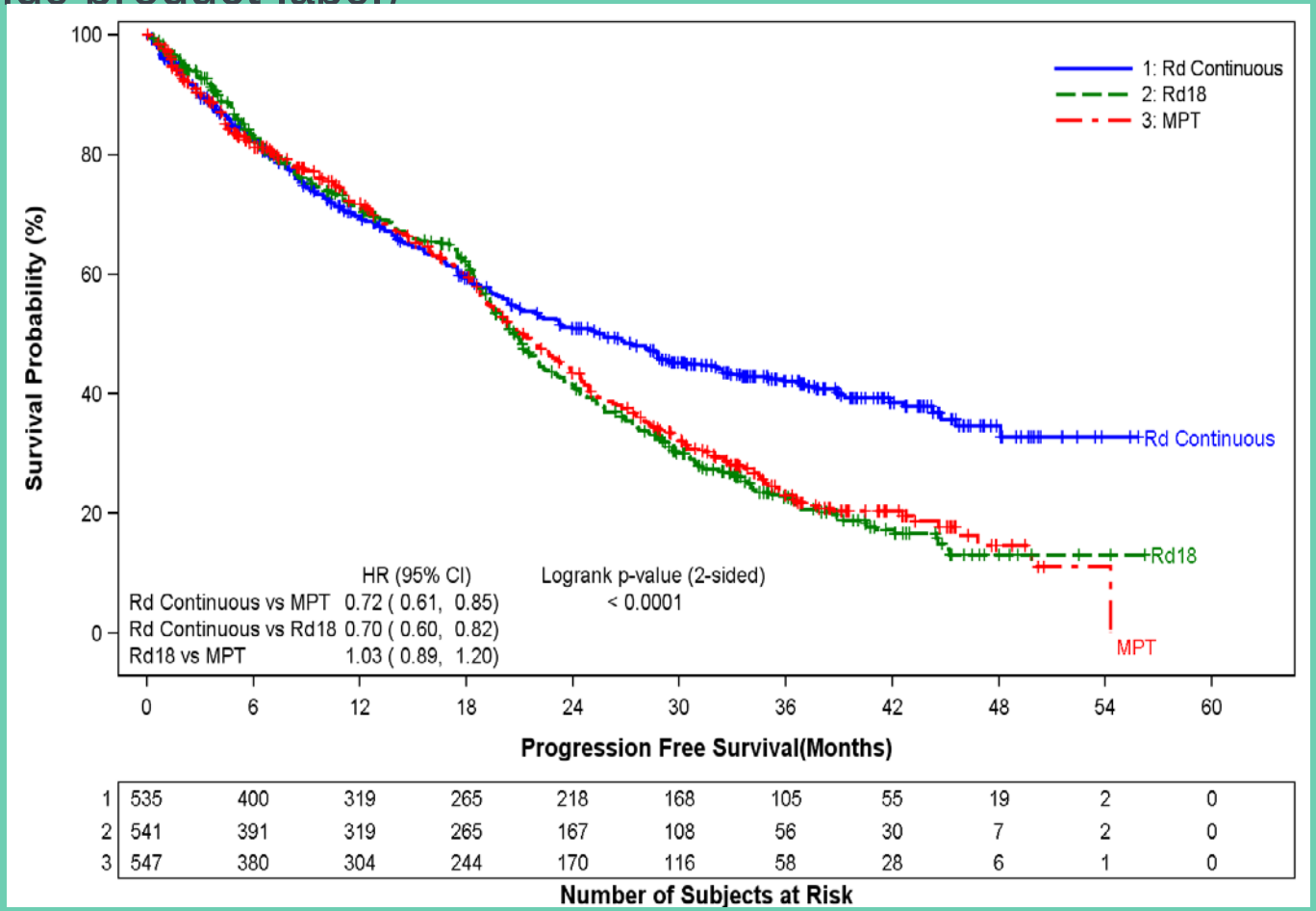
- Endpoint Examples: Overall survival (OS), Progression-free Survival (PFS), Recurrence-free Survival (RFS), Disease-free Survival (DFS), etc.
- Most common approach in design of clinical trials with time to event endpoint:
  - Fix chance of false positive conclusion (alpha)
  - Fix chance of winning or detecting benefit if it exists (power of the test)
  - Define what treatment effect is meaningful (alternative hypothesis to null hypothesis of no effect)
  - Assume relative treatment effect (hazard ratio) is **constant** over time

Standard Time to Event Analysis

Assuming constant relative treatment effect over time,

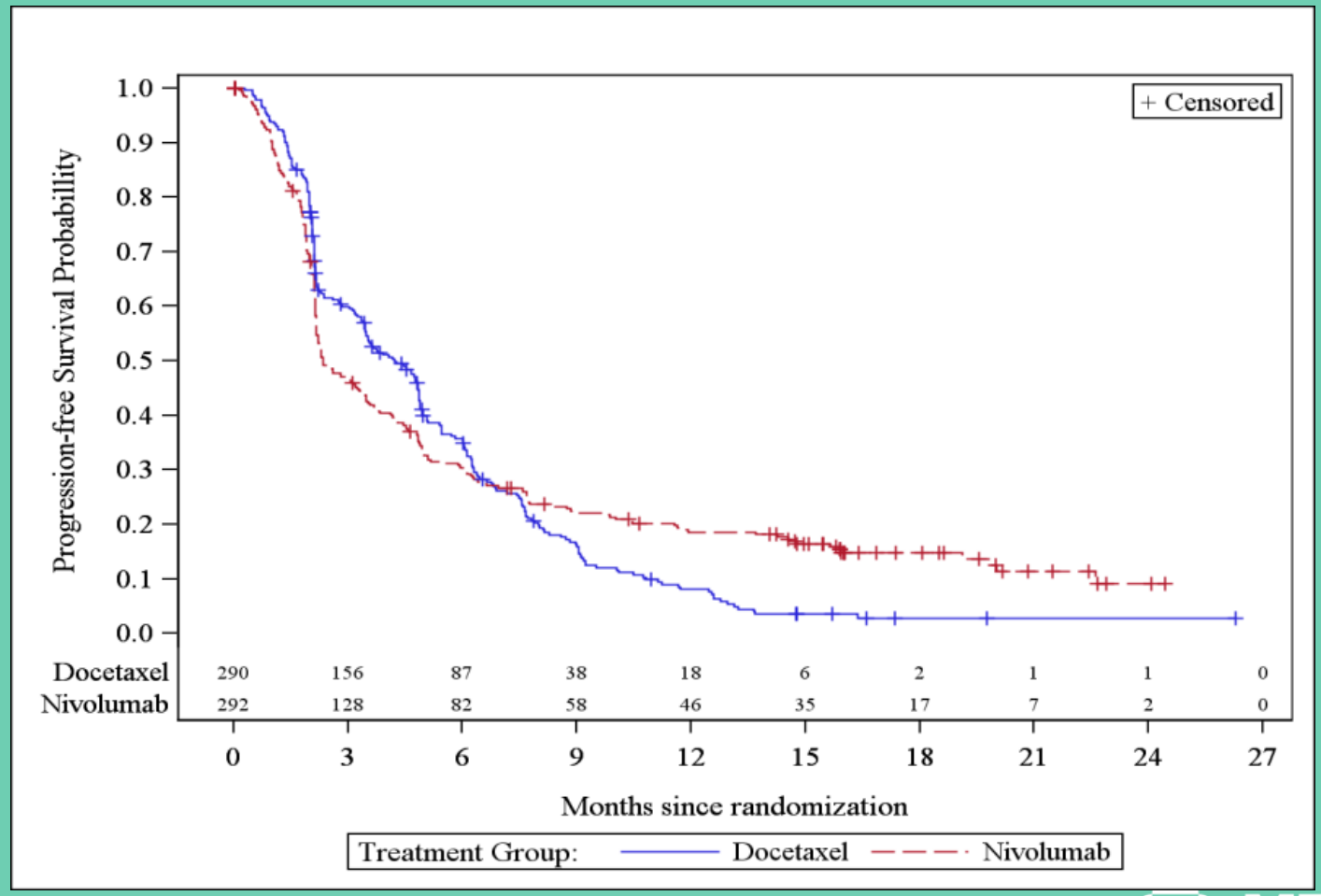
- Comparison of survival curves using Log-rank test (Non-parametric test)
  - Estimated median survival provides a summary of the survival curve (i.e., on an average, 50% of events observed before the median time)
- Test hypothesis and estimate relative treatment effect using Cox-proportional hazards model
  - Hazard ratio provides an average relative effect over time
- Power to test the hypothesis reduces as relative effect changes over time (violation of the constant effect assumption)

Kaplan-Meier Curves of Progression-free Survival Based on IRAC Assessment (ITT Population) Between Arms Rd Continuous, Rd18 and MPT (Lenalidomide product label)



MERCK

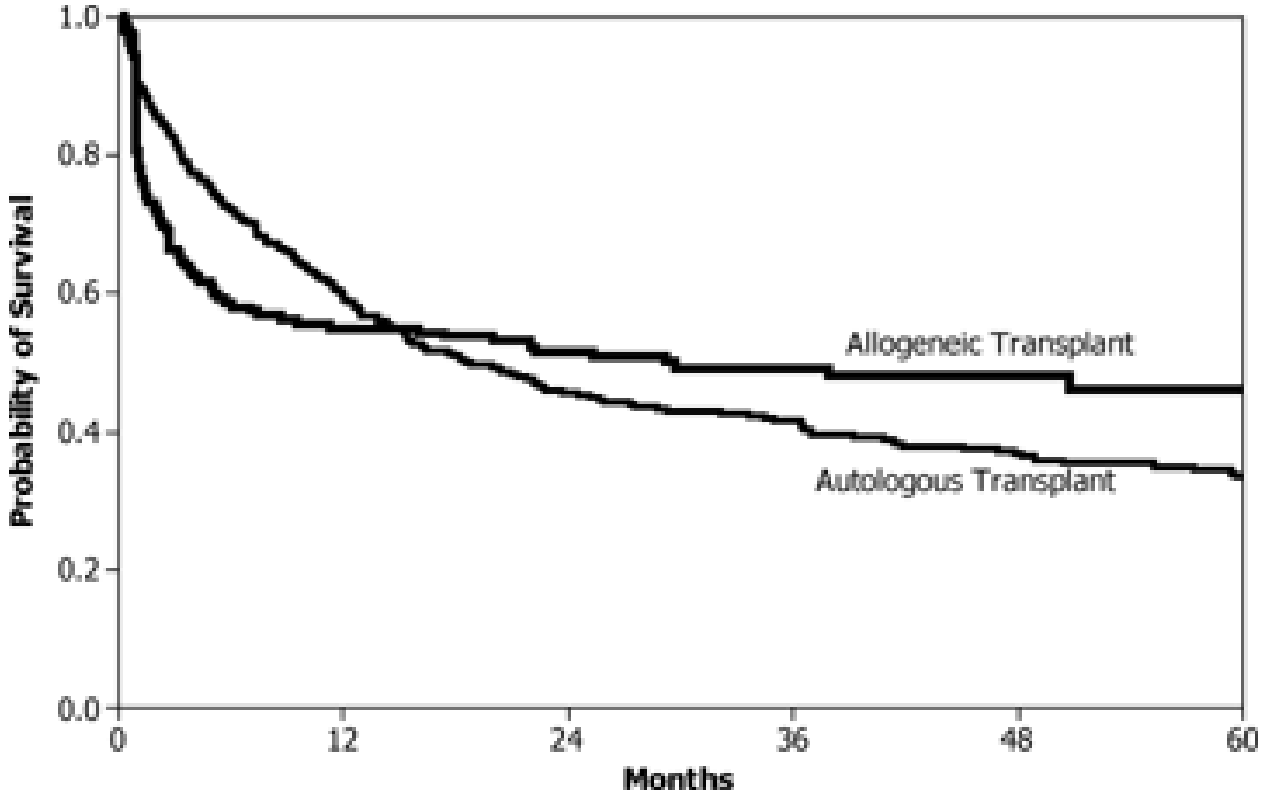
Nivolumab 2nd line non squamous mNSCLC: PFS analysis



Sridhara DIA 6-21-17

MERCK

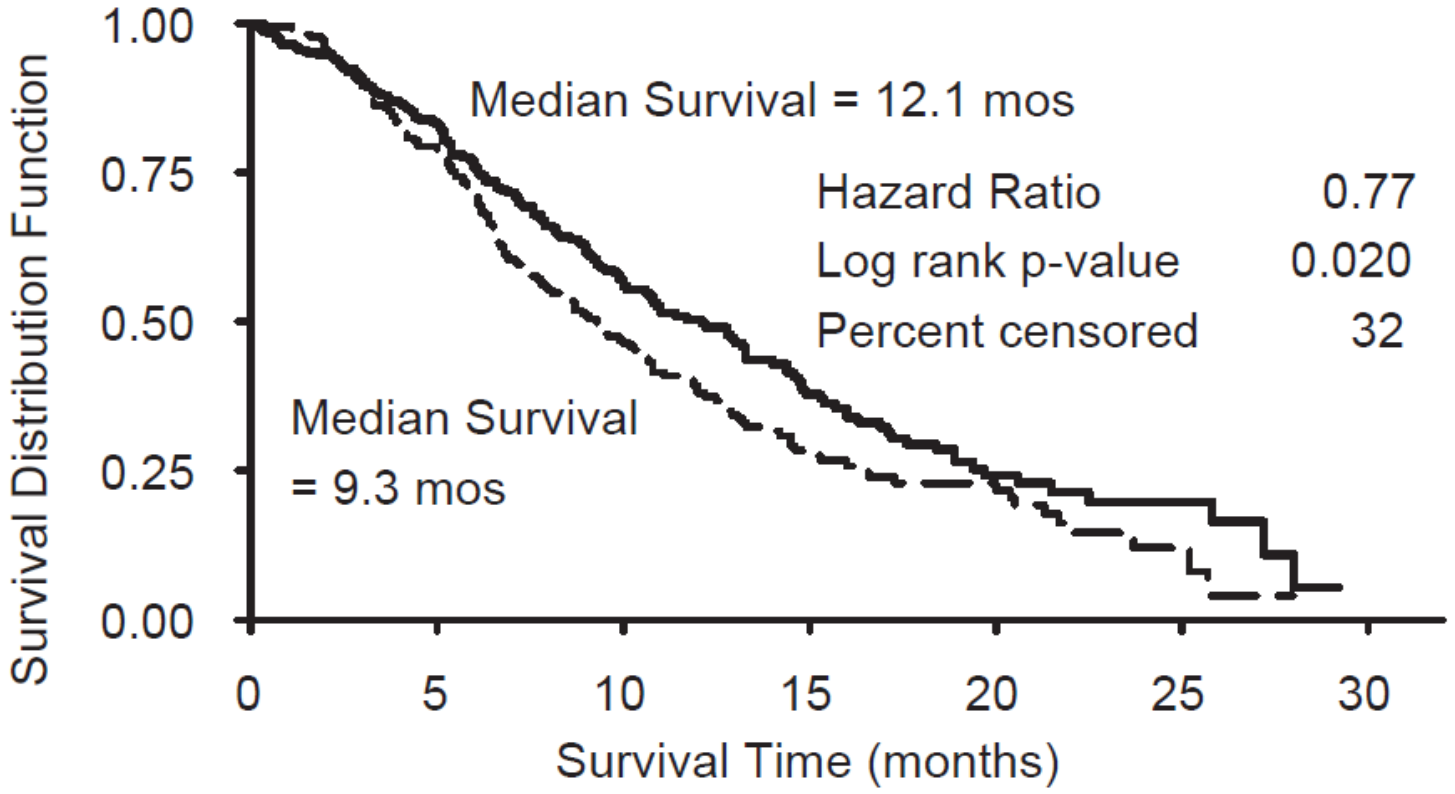
Comparing Treatments in the Presence of Crossing Survival Curves: An Application to Bone Marrow Transplantation



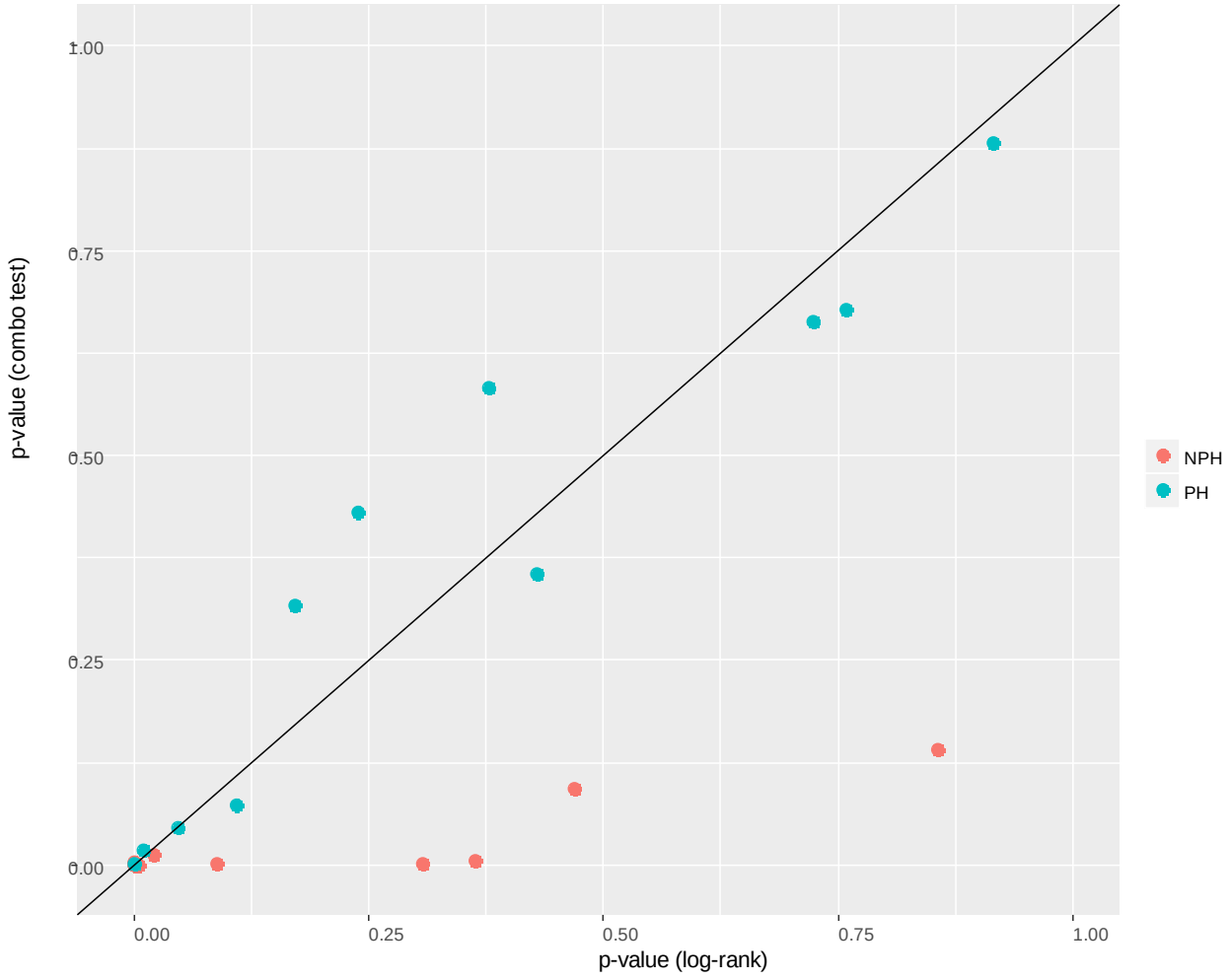
Kaplan-Meier estimate of DFS for Follicular Lymphoma by transplant source

Biometrics
Volume 64, Issue 3, pages 733-740, 11 JAN 2008 DOI: 10.1111/j.1541-0420.2007.00975.x
http://onlinelibrary.wiley.com/doi/10.1111/j.1541-0420.2007.00975.x/full#f1

Pemetrexed in Mesothelioma (product label)



P-values: Max-combo test *vs.* log-rank test



- The use of weights in the max-combo test suggests that some events are “more important” than others. How to justify it?
- Source: Lijun Zhangj, FDA Duke-Margolis slide set

Challenges

- When the assumption of constant HR is not true,
  - Cox-proportional hazard model is inappropriate
  - KM estimate of median survival may not be an optimal measure to summarize the results
- What is an optimal analysis method to test treatment benefit and how can we summarize the benefit?
  - Many methods have been proposed in literature; each have advantages and limitations
  - Multiple approaches may be necessary to summarize results

FDA Initiated Collaboration

- FDA recognized the need for collaboration
- Initiated dialogue with with the Industry statisticians
- Met in 2016 and subsequently in 2017
  - Concluded that a methodical evaluation of available methods is needed
  - Goal: identify appropriate analysis method for the different patterns of non-proportionality
  - All industries to work together as a team (non-product specific)
  - FDA to participate in this effort

Why are we here today?

- Current practice of using log-rank test and Cox-proportional hazards model not appropriate when relative treatment effect varies over time (Non-proportional Hazards)
- Reasons for observed changes in treatment effect over time may be different in different clinical trials
- What is the best way to evaluate treatment effect?
- What is the best way to summarize an observed treatment effect?
- Working group will be presenting what has been accomplished so far

ANALYSIS METHODS

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

- Most popular methods in randomized clinical trial:
  - Kaplan-Meier (KM): describe chance 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 shows evidence of a delayed effect
- How to cope with NPH problem at design and analysis stages?

MERCK

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
  - **asymptotically efficient** for proportional hazards (PH)
  - **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 is violated

MERCK

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?

MERCK

Choice of Primary Analysis in Confirmatory Trials

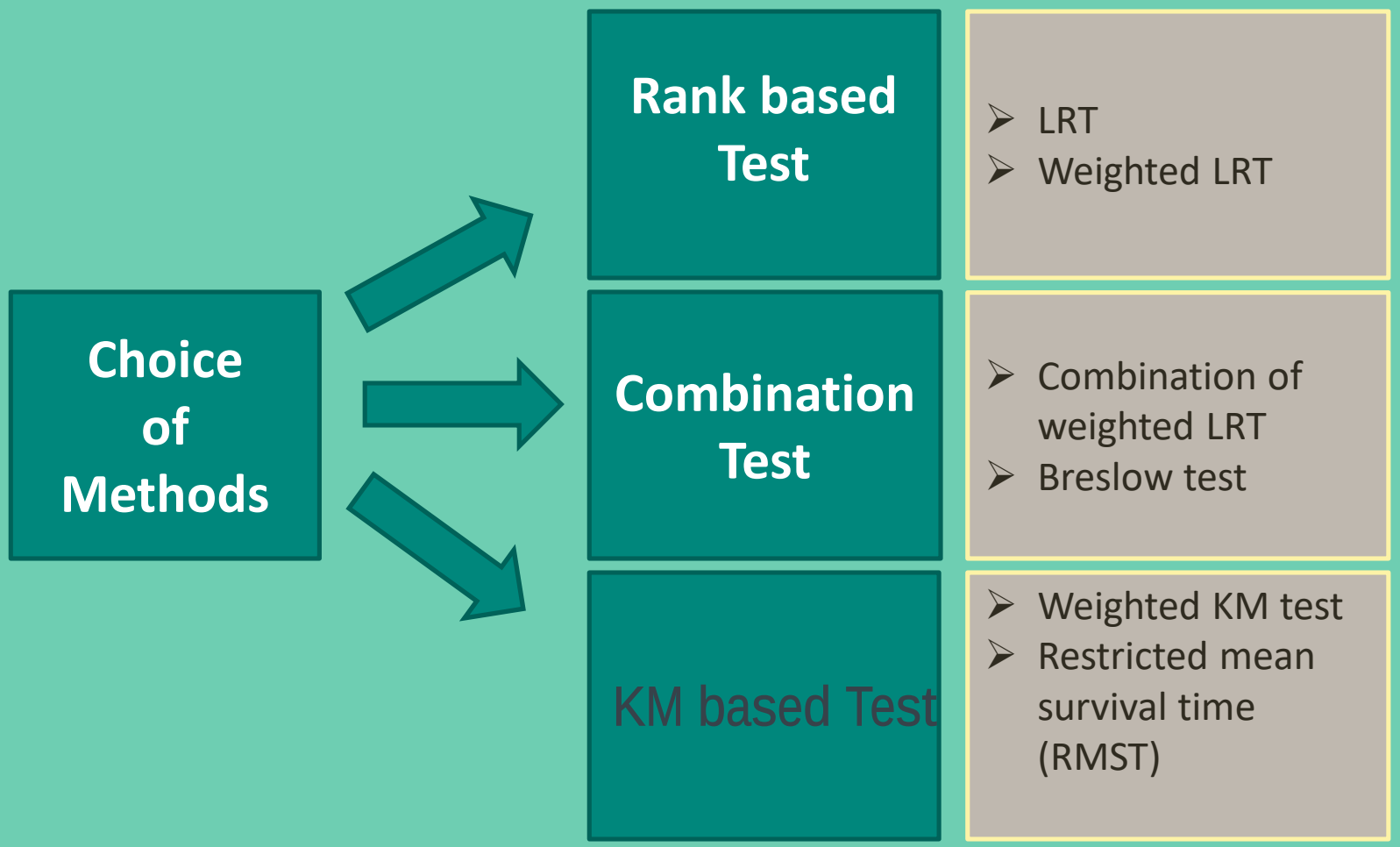
- Regarding **primary analysis** ICH E9 states

  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)

MERCK

Choice of Primary Analysis Methods



Weighted Log-rank Test

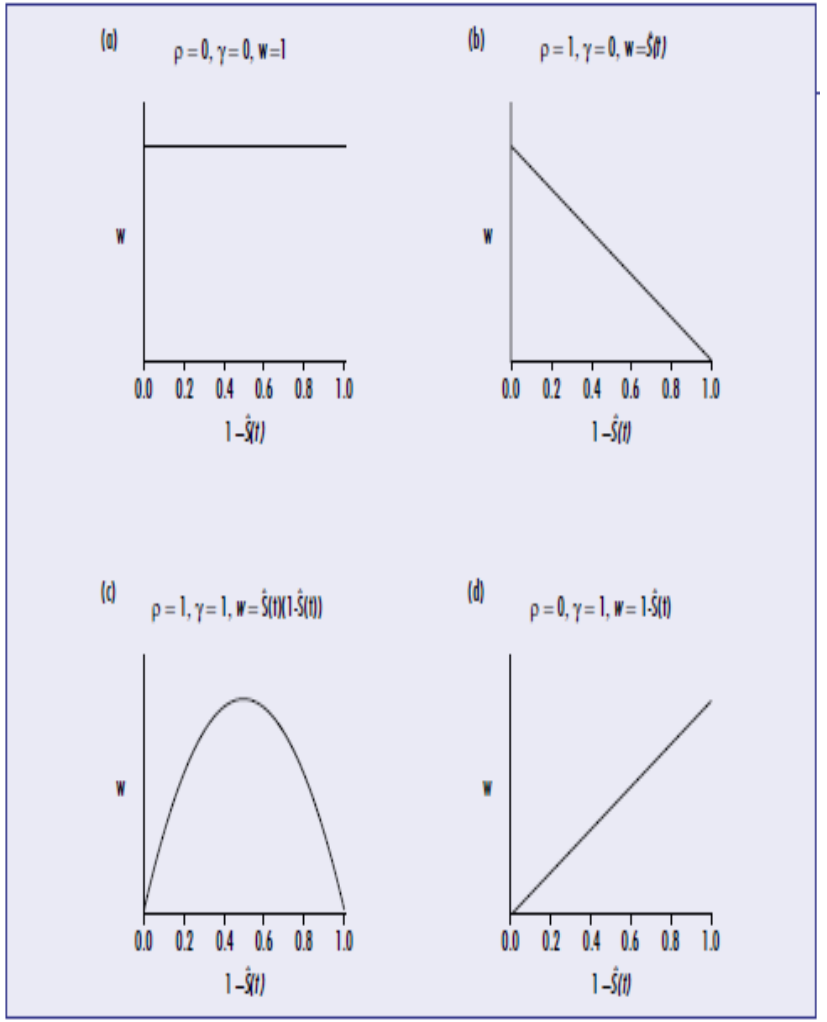
Fleming and Harrington proposed a class of weighted log-rank test (FH) based on the $G^{\rho,\gamma}$ family

Assign weight to events

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

Values of ρ and γ implies

- ρ > 0, γ =0 : early difference
- ρ =0, γ >0 : late difference
- ρ >0, γ >0 : mid difference
- ρ =0, γ =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*
  - Karrison 2016: *Versatile tests*
  - Garès et. al. 2017: maximal statistics over FH(0,γ)

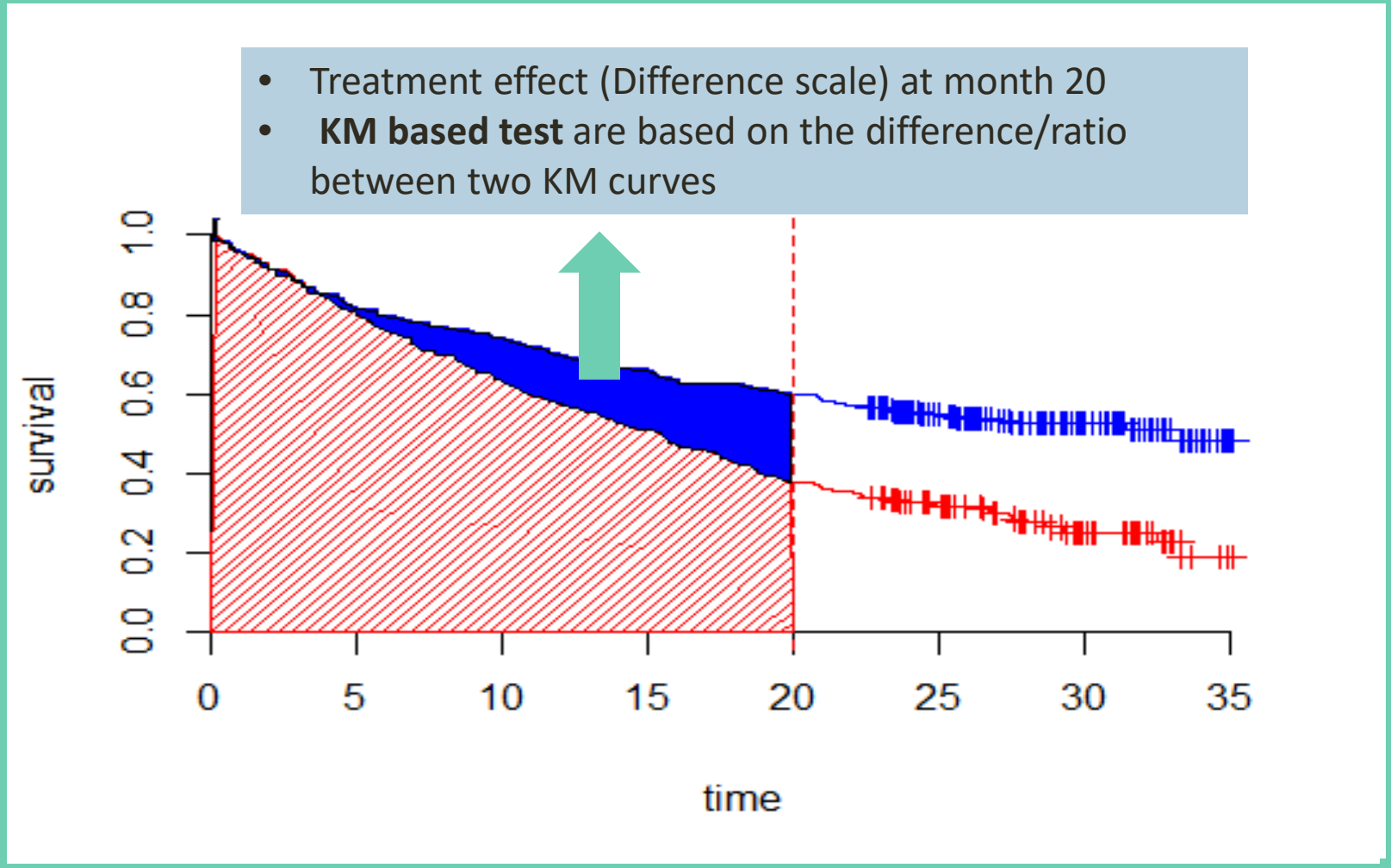
MERCK

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-Holm 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**

MERCK

Kaplan-Meier Based Tests



Data cutoff

- Take maximum follow-up in each treatment group
- Minimum of these maxima is cutoff
- Recent justification for this for RMST submitted for publication

MERCK

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 expectancy"
  - treatment effect: difference or ratio of RMST
  - need to specify the cut-off: can be affected by censoring

MERCK

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
- Net chance of longer survival: Buyse (2010), Peron et al (2018)
  - Generalized pairwise comparison
  - Can specify ‘clinically significant’ difference for pairwise evaluations

MERCK

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*
- *Net chance of longer survival*
  - *Interesting with pre-specified cutoff or as a function of minimum important difference?*

MERCK

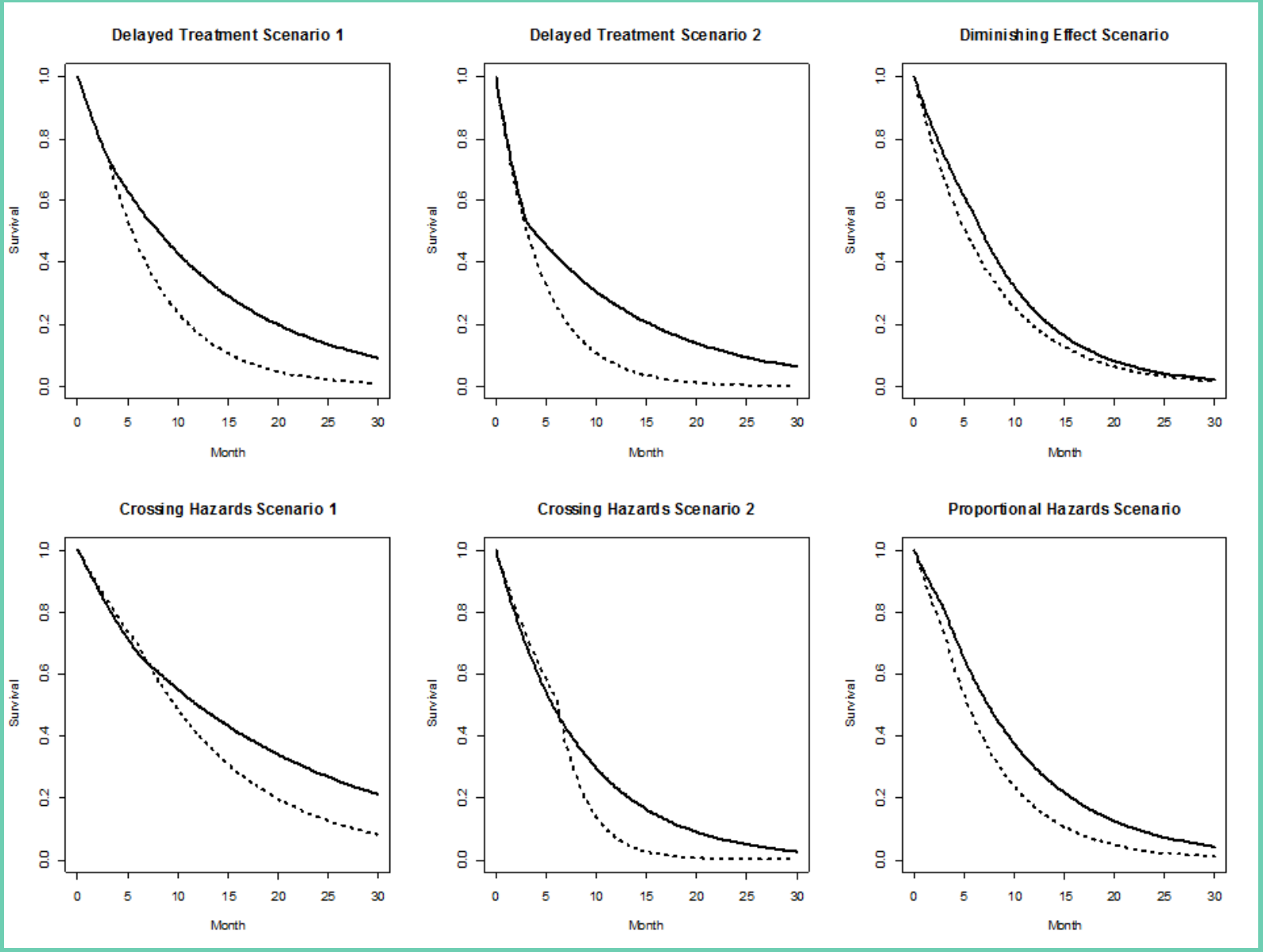
Choice of Treatment Effect Summary

- 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** and **average** HR (**Max-Combo**) with 95% CI
  - **difference in RMST at max cutoff**
  - **difference in milestone survival at t^***: gain in chance of survival at clinically relevant time point t^* (pre-specified)
  - secondary analysis: piecewise HR and/or piecewise failure rates with 95% CI

MERCK

SIMULATION STUDIES

Simulation scenarios studied



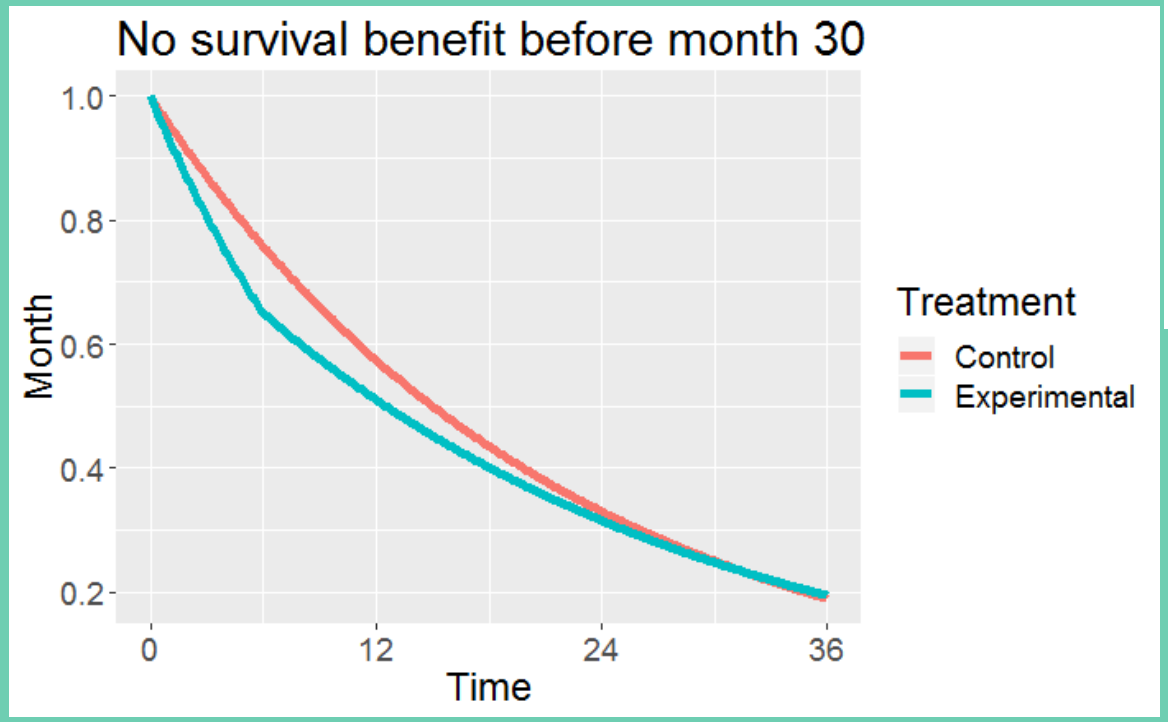
Simulation results

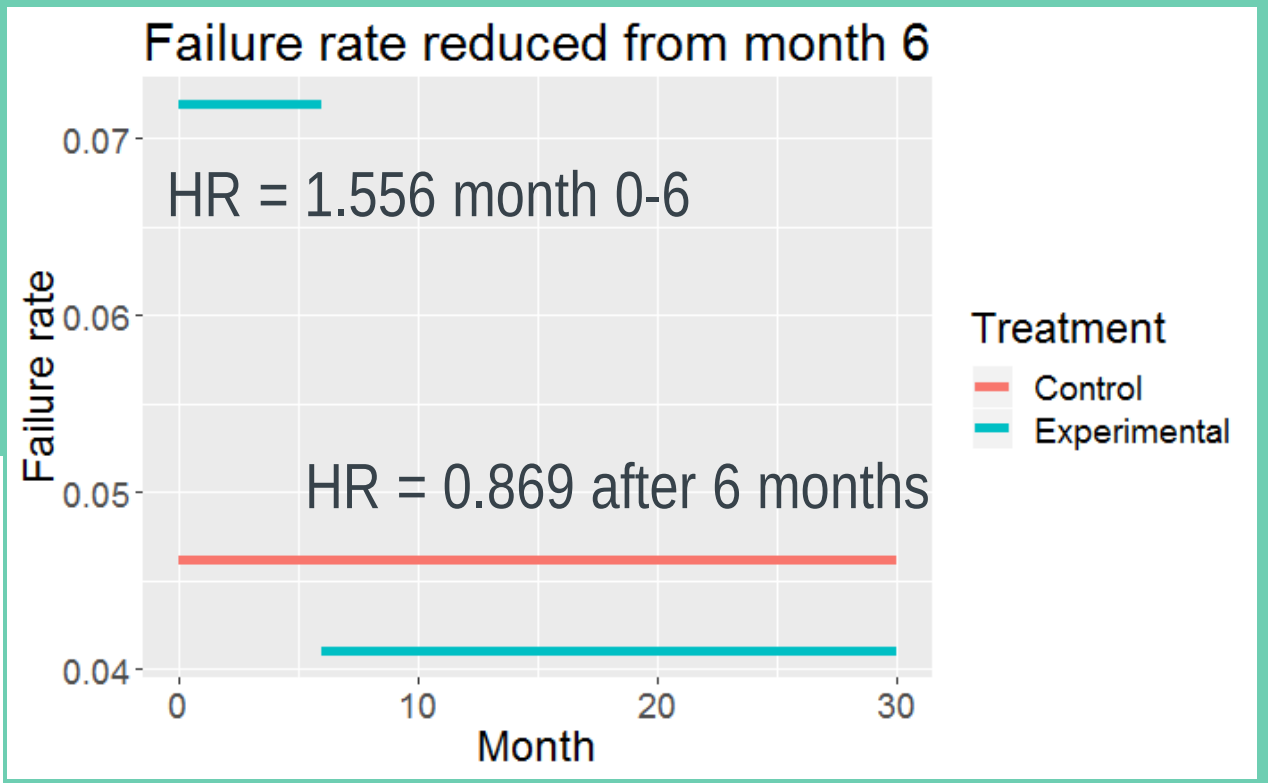
- MaxCombo had competitive power for all scenarios
- Type I error controlled when survival is equal
- Individual tests performed poorly in at least some scenarios

MERCK

What is null hypothesis space for weighted logrank?

For weighted logrank, benefit measured as a function of relative failure rates





- This may not correspond to a survival benefit
- For increasing weights, this can be out of null hypothesis space

MERCK

Type I error controlled by MaxCombo?

- Underlying survival distribution
  - Controls exponential with median of 15 months (λ=0.046)
  - Experimental group is piecewise exponential
    - HR=1.556 for 6 months
    - HR=0.869 thereafter
    - Survival curves cross at 30 months
- Enrollment: N=200
  - Constant enrollment rate for 12 months
- Data cutoff: 30 months
- Type I error (1-sided; 10k simulations)
  - ✓ MaxCombo: 1.5%
  - ✓ MaxCombo also requiring upper CI for HR < 1.1: 0.78%
  - ❖ Inflated for FH(0,1): 2.7% (within simulation error)

❖ There are potential issues here in some cases

MERCK

STUDY DESIGN

Design issues

- Trials results often differ from design assumptions
- Results may differ by
  - Degree of effect
  - Delayed timing of effect
    - Delayed separation of survival curves
  - Different effects in unanticipated subpopulations
    - This can result in crossing hazards
  - Diminishing effect over time
    - Converging hazards – maybe of LESS interest here
- How do we design a trial to be powerful across MANY alternatives?

MERCK

Design philosophy

- Power trial for multiple scenarios
- Find worst-case scenario, e.g.,
  - Minimum effect size of interest (PH)
  - Delayed effect
  - Early crossing hazards
- Simple approximation of alternatives
  - Piecewise exponential failure
  - Single change point
- No single estimand/estimate is adequate
  - Inconsistent with ICH E9 (R2) estimand recommendations?

MERCK

Design implementation

- Ensure adequate follow-up
- Robust testing method
- If using MaxCombo
  - Karrison (2016) provides correlations needed to adjust for multiple tests
  - Power for multiple scenarios & select worst-case sample size
    - Use adjusted significance level for components of MaxCombo
      - Modification of Hasegawa (2016) for calculation
      - Power for best MaxCombo component will be conservative

MERCK

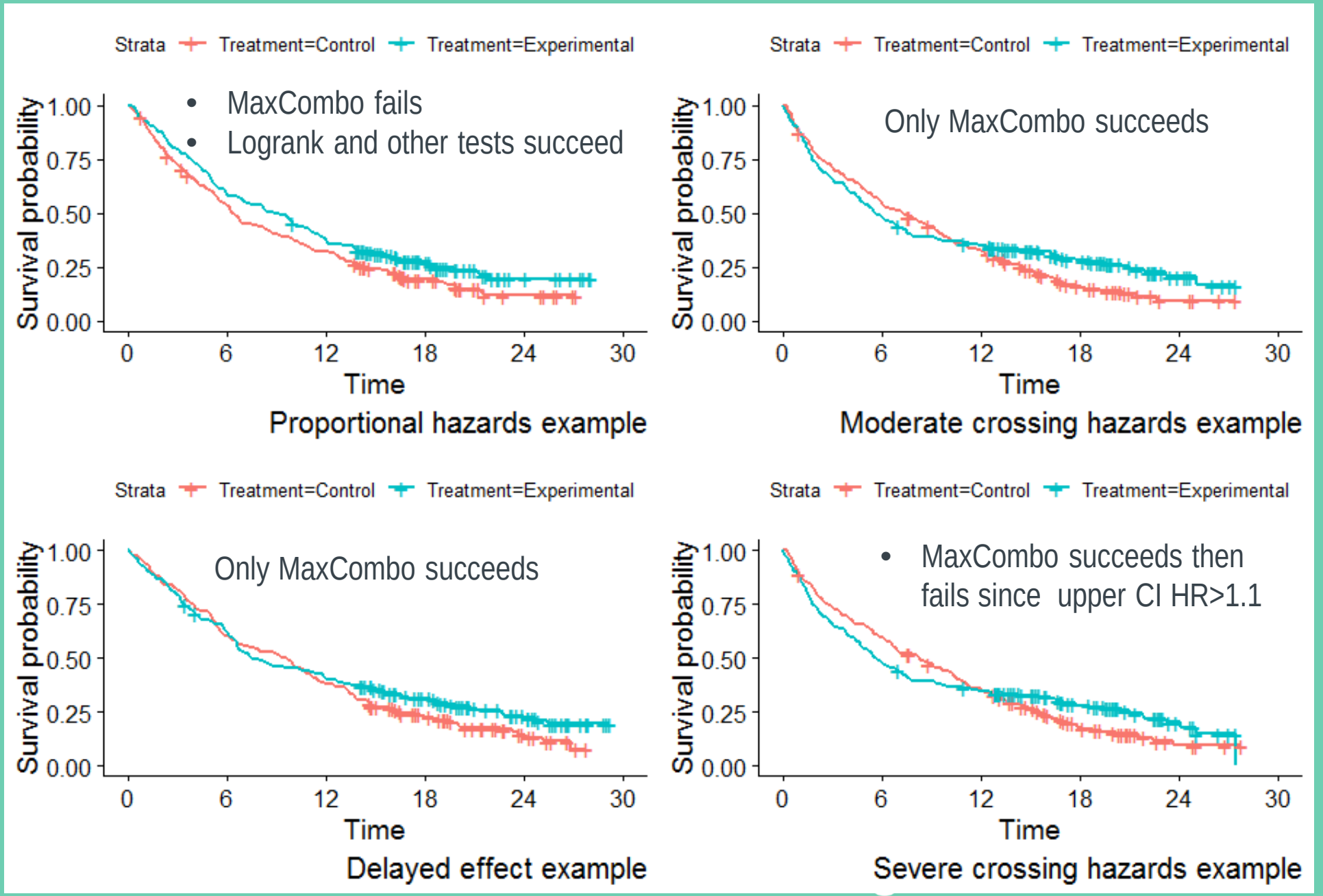
Design: interim analysis (IA) considerations

- Recommend logrank for interim stopping
  - Improve regulatory acceptance?
  - May wish to use MaxCombo for sensitivity analysis
- Lack of efficacy
  - Are early tests of excess mortality required?
    - Early safety bounds rather than futility bounds
  - Conditional power-based futility: Freidlin and Korn (201?)
- Efficacy testing
  - Delayed effect may result in fast event accumulation
  - Set timing based on events AND follow-up to ensure power

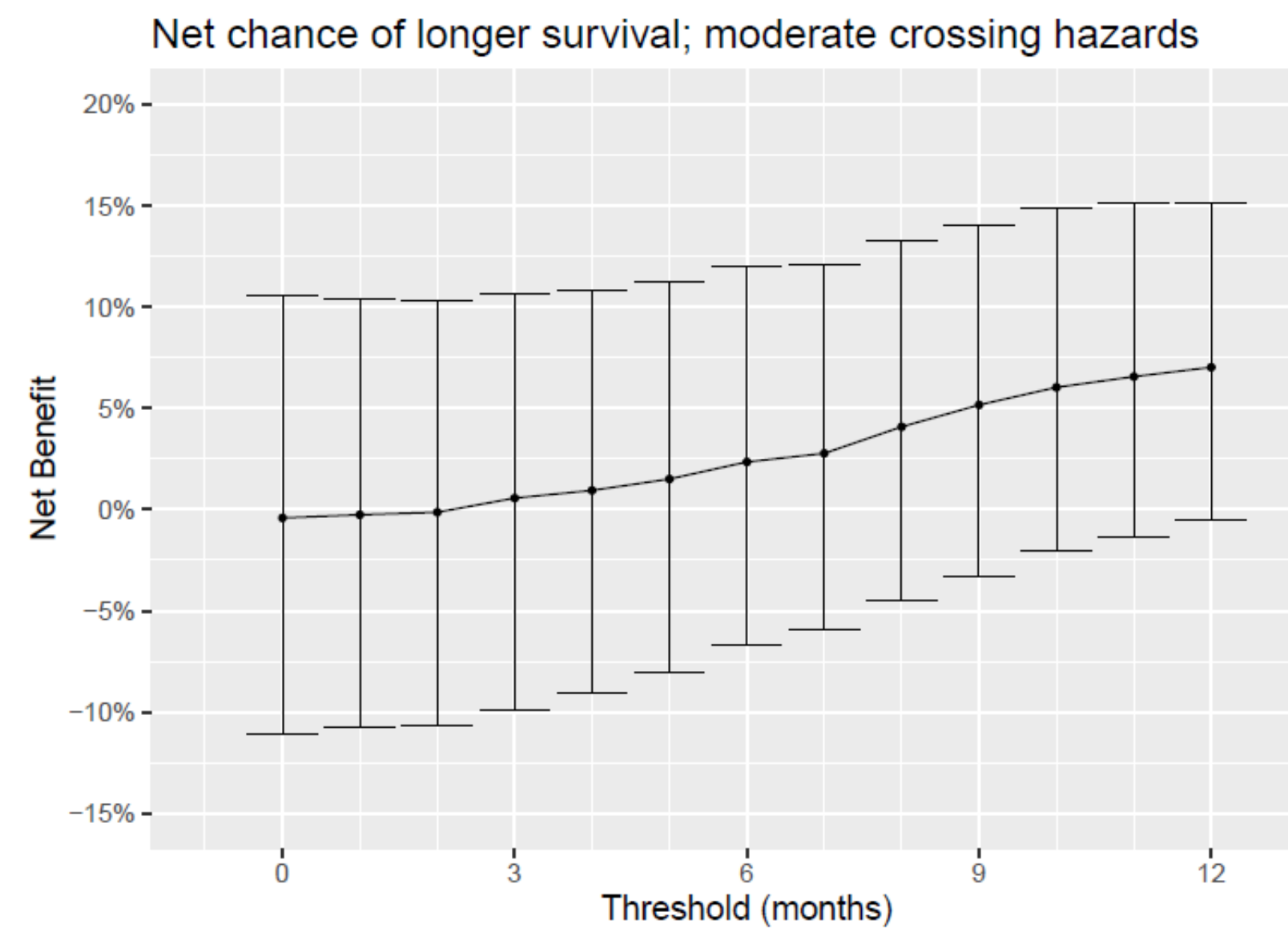
MERCK

BREAKDOWN AND ESTIMATION EXAMPLES

Breakdown examples



Net chance of longer survival example



- Preferred cutoff may be patient-dependent
- Power not well-studied in our simulations
  - For examples, was not positive other than for PH
- Is this helpful beyond Kaplan-Meier curve?

MERCK

Summarizing benefit
Moderate crossing hazards example

Analysis	Experimental	Control	Estimate (95% CI)	p-value
Median/HR/logrank	5.594	7.303	0.878 (0.708,1.089)	0.118
Weighted HR/MaxCombo	NA	NA	0.689 (0.515,0.923)	0.004
RMST	10.544	9.503	1.041 (-0.767,2.849)	0.130
RMTL	16.941	17.982	0.942 (0.849,1.046)	0.131
% favorable by 6 mos	25.244	26.297	-1.054 (-10.415,8.056)	0.593
Weighted KM	NA	NA	NA	0.367

MERCK

Summarizing benefit: Milestone survival
Moderate crossing hazards example

Month	Experimental	Control	Difference	(95% CI)
3	66%	71.5%	-5.5%	(-14.2%,3.3%)
6	48.4%	55.2%	-6.8%	(-16.2%,2.6%)
12	34.8%	32.8%	2%	(-7%,11%)
18	27.9%	16.1%	11.8%	(3.6%,20.1%)
24	20.6%	9.4%	11.2%	(2.1%,20.3%)

MERCK

Summarizing benefit: Piecewise exponential failure rates
Moderate crossing hazards example

Period	Experimental	Control	HR (95% CI)
0-3 months	0.139	0.113	1.237 (0.88,1.737)
3-6 months	0.103	0.086	1.194 (0.754,1.89)
6-12 months	0.058	0.087	0.666 (0.419,1.057)
>12 months	0.038	0.109	0.350 (0.199,0.616)

SUMMARY

Potential concerns for alternative methods for regulatory approval

- Focus here on metastatic (high-risk) scenario
  - Long-term outcomes with low rates may require alternate approach
- Proposed estimand for MaxCombo not intuitive
  - Weighted HR based on best FH weighting
  - Descriptive alternatives
    - Milestones, piecewise rates and piecewise HR
- Type I error for theoretical cases with no benefit
  - Sponsor needs to justify Type I error protection
  - FURTHER CLARIFICATION NEEDED.
- Primary concern was delayed treatment effect
  - Alternatives other than weighted approaches not doing well?

MERCK

Where is the NPH working group now?

- Near-final draft of simulation paper
- Draft paper on design and analysis prepared
- Estimand working group now working in parallel
- Need for further regulatory interaction

MERCK

Conclusions

- MaxCombo useful for non-proportional hazards in metastatic setting
- Important benefit could be missed with other methods
- Proposals are ready for alternatives to logrank/Cox/median
- Sponsors encouraged to submit as supportive
- Further discussion needed to move approaches to primary

References

1. Breslow, N et. al. (1984). A two sample censored data rank test for acceleration. Biometrics, 40: 1042–1069

2. Buyse, Marc. 2010. “Generalized Pairwise Comparisons of Prioritized Outcomes in the Two-Sample Problem.” *Statistics in Medicine* 29 (30). Wiley Online Library: 3245–57.

3. Cox DR (1972). Regression-Models and Life-Tables. Journal of the Royal Statistical Society. B (Methodological), 34 (2): 187–220

4. Fleming TR, Harrington DP (1991) Counting Processes and Survival Analysis. John Wiley & Sons: New York

5. 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

6. Grambsch, Patricia M, and Terry M Therneau. 1994. “Proportional Hazards Tests and Diagnostics Based on Weighted Residuals.” *Biometrika* 81 (3). Oxford University Press: 515–26.

7. Hasegawa, Takahiro. Group sequential monitoring based on the weighted log-rank test statistic with the fleming–harrington class of weights in cancer vaccine studies. Pharmaceutical Statistics, 15(5):412–419, 2016.

8. Karrison, Theodore G, and others. 2016. “Versatile Tests for Comparing Survival Curves Based on Weighted Log-Rank Statistics.” *Stata Journal* 16 (3). StataCorp LP: 678–90.

9. Pepe, Margaret Sullivan, and Thomas R Fleming. 1989. “Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data.” *Biometrics*. JSTOR, 497–507.

10. ———. 1991. “Weighted Kaplan-Meier Statistics: Large Sample and Optimality Considerations.” *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 341–52.

11. Péron, Julien, Marc Buyse, Brice Ozenne, Laurent Roche, and Pascal Roy. 2018. “An Extension of Generalized Pairwise Comparisons for Prioritized Outcomes in the Presence of Censoring.” *Statistical Methods in Medical Research* 27 (4). SAGE Publications Sage UK: London, England: 1230–9.

12. Péron, Julien, Pascal Roy, Brice Ozenne, Laurent Roche, and Marc Buyse. 2016a. “The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials.” *JAMA Oncology* 2(7). American Medical Association: 901–5.

13. ———. 2016b. “The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials.” *JAMA Oncology* 2 (7). American Medical Association: 901–5.

14. Royston, Patrick, and Mahesh KB Parmar. 2011. “The Use of Restricted Mean Survival Time to Estimate the Treatment Effect in Randomized Clinical Trials When the Proportional Hazards Assumption Is in Doubt.” *Statistics in Medicine* 30 (19). Wiley Online Library: 2409–21.

15. Uno, Hajime, B. Claggett, L. Tian, and et. al. 2014. “Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis.” *Journal of Clinical Oncology*.

THANK YOU

MERCK