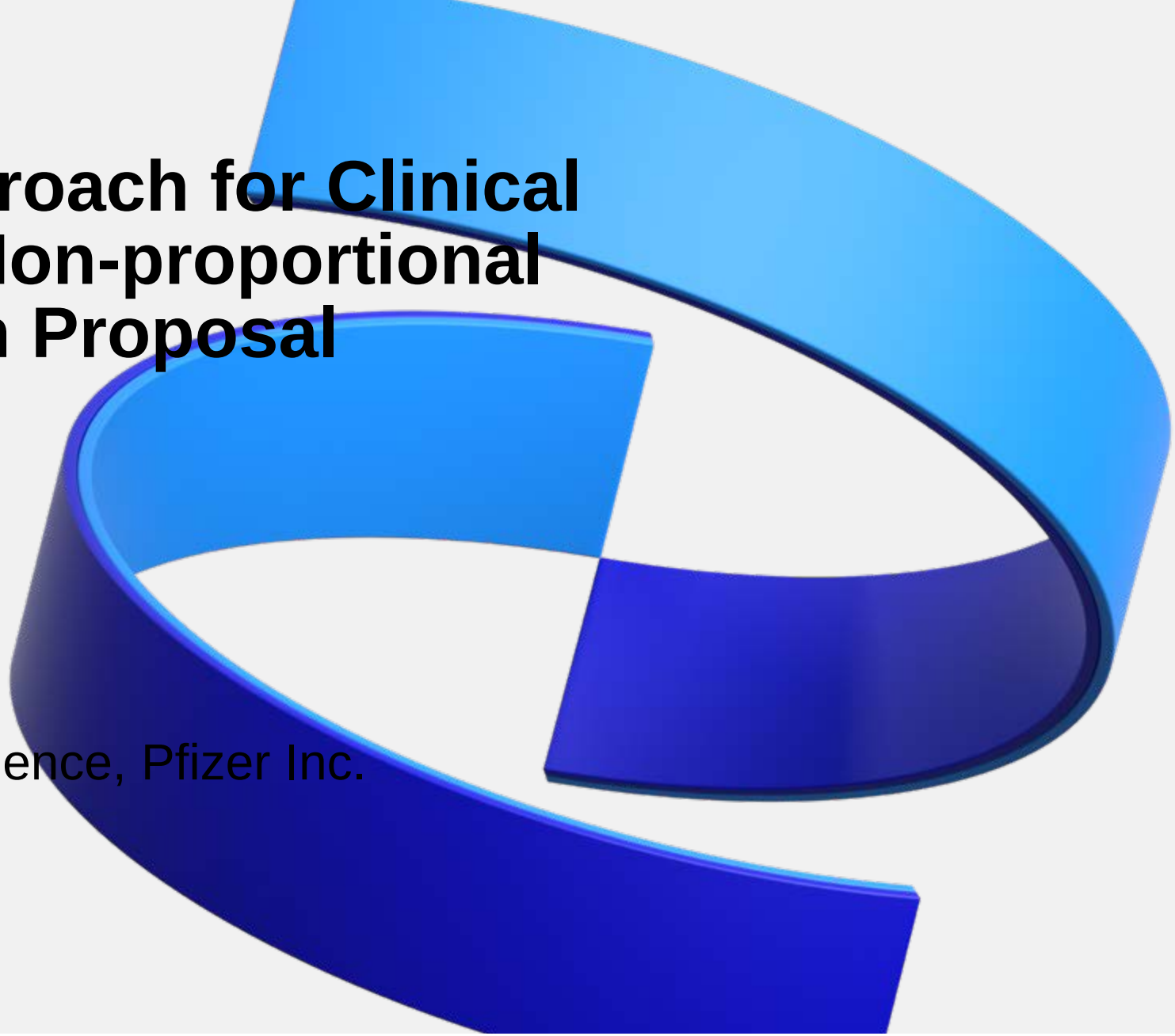
A Robust Design Approach for Clinical Trials with Potential Non-proportional Hazards: A Straw Man Proposal



Satrajit Roychoudhury
Statistical Research and Data Science, Pfizer Inc.

PSI Scientific Committee Meeting
April 29, 2021

Disclaimer

While the author is a member of the Non-Proportional Hazards (NPH) Working Group, any mistakes and opinions should be considered those of the author. Also, this work does not represent a company position for Pfizer Inc.

Acknowledgement

- **Joint work with**
  - o Keaven Anderson (Merck & Co.)
  - o Ji Lin (Sanofi)
  - o Ray Lin (Roche)
  - o Pralay Mukhopadhyay (Otsuka Pharmaceuticals)
  - o Jaibu Ye (Merck & Co.)
  - o Renee B Iacona (Astrazeneca Pharmaceuticals)
  - o Tai-Tsang Chen (Glaxosmithkline)
  - o Members the of cross-pharma non-proportional hazards working group

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

Treatment effect not constant over time

Most popular analysis methods in randomized clinical trial with time to event endpoint:

- Kaplan-Meier (KM): describe chance of survival over time
- log-rank test (LRT): detect difference in treatment effect (rejects “Null”)
- Cox regression (CR): summarize the treatment effect

Log-rank p-value, hazard ratio, and KM medians are the standard metrics of reporting

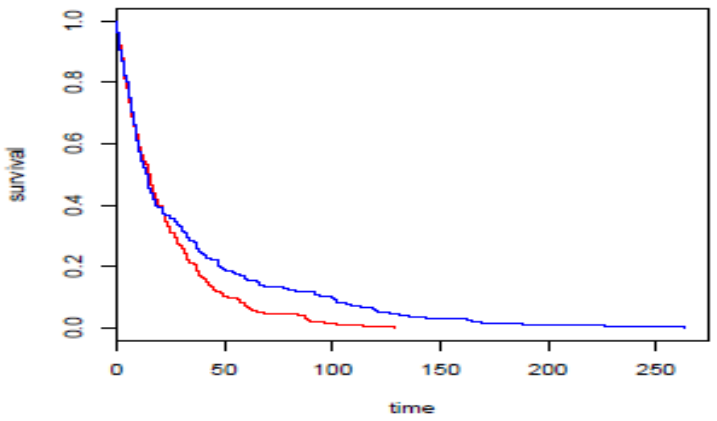
Are these good summary measures when the treatment effect is not constant over time?

- For example, recent immunotherapy development shows evidence of a delayed effect

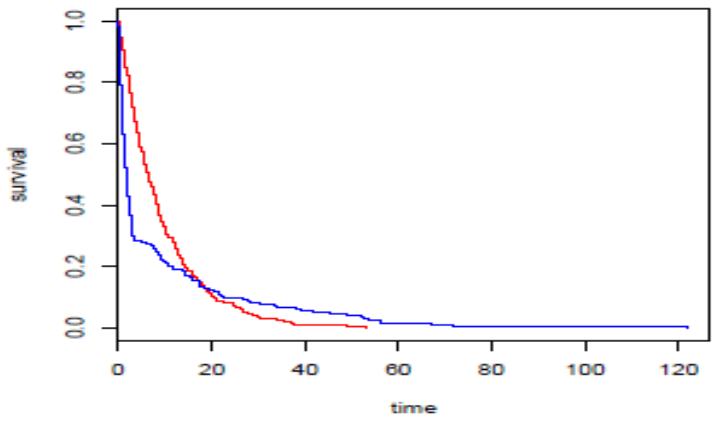
How to cope with NPH problem at design and analysis stages?

Different Types of NPH

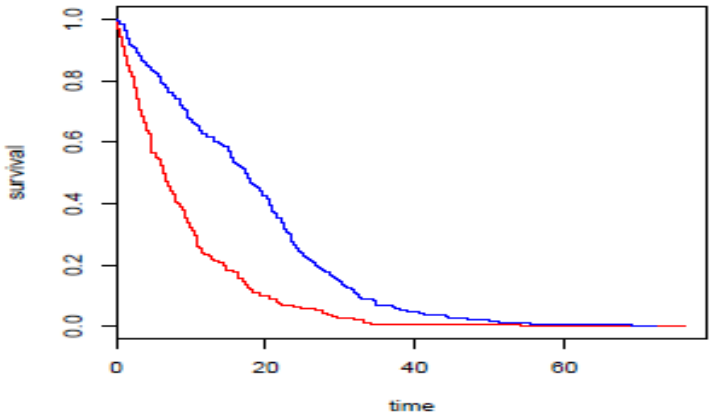
Delayed Treatment Effect



Crossing Survival

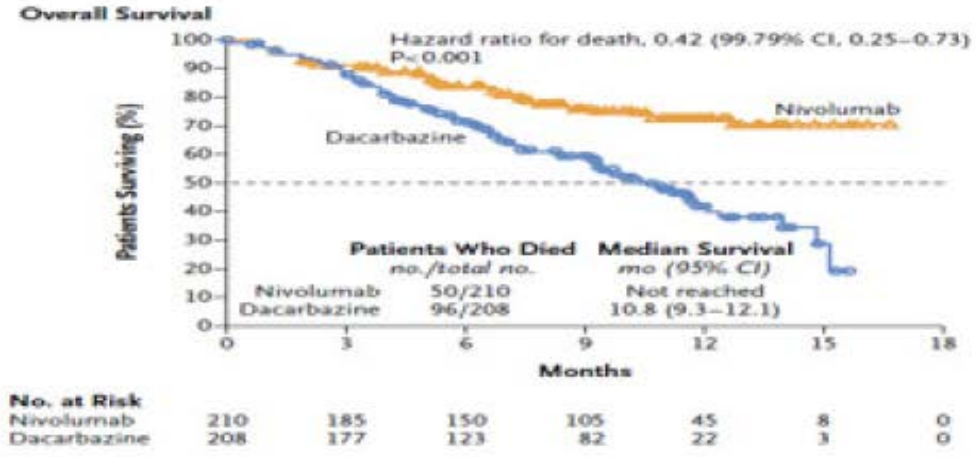


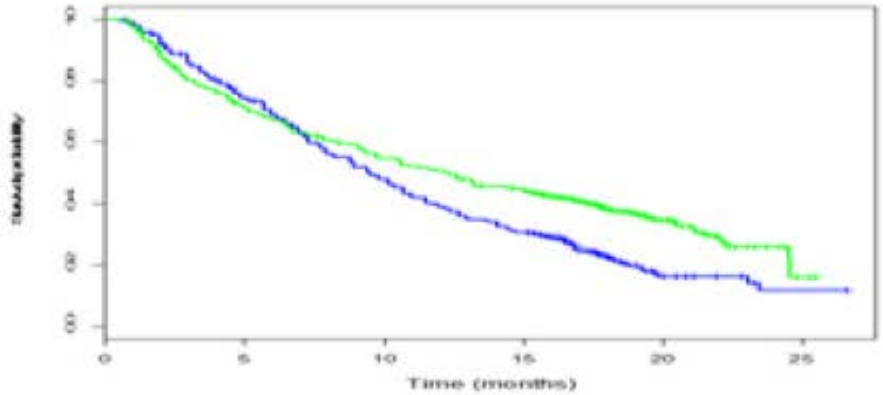
Diminishing Treatment Effect

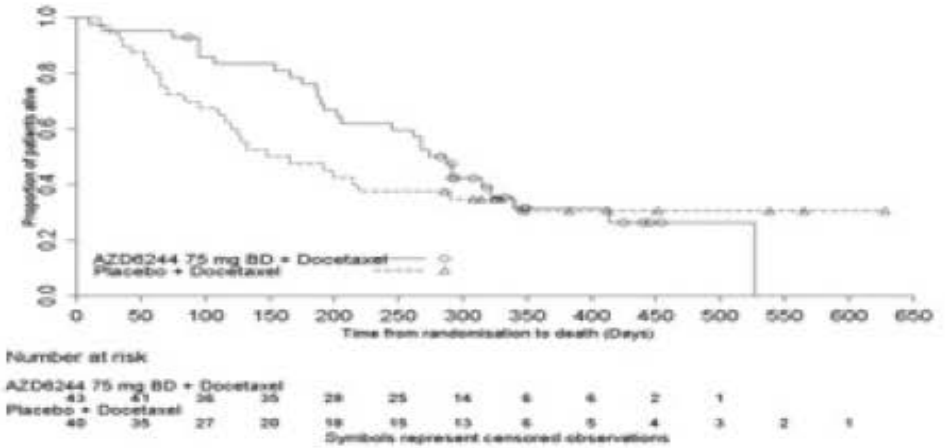


- Unlike PH, NPH is not unique
- Uncertainty related to the type of NPH when trial starts

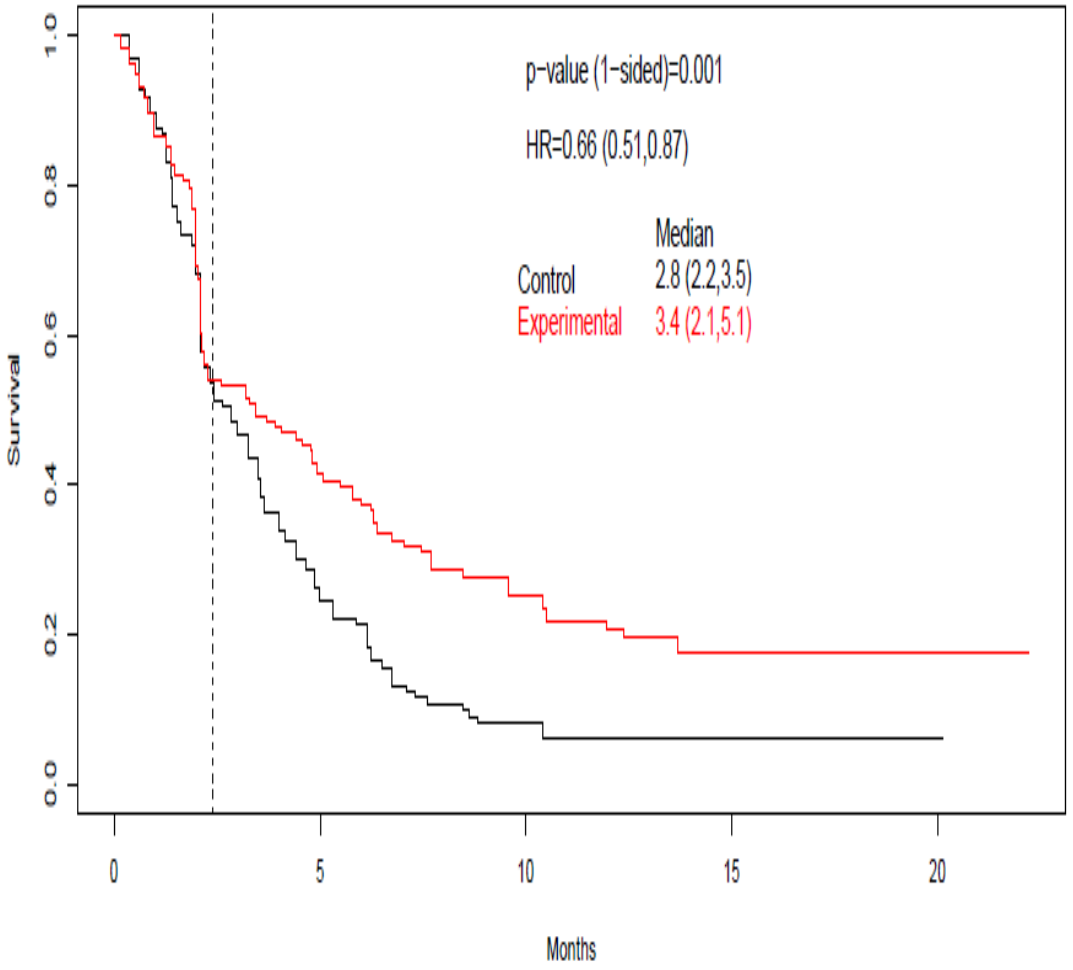
Recent Examples from Oncology Trials







Treatment Effect Emerges Late in the Trial



Information Fraction	No. of Events	Time (month)	HR	95% CI
22%	49	1.4	0.906	(0.52, 1.59)
49%	110	2.1	0.933	(0.64, 1.36)
52%	118	2.2	0.971	(0.68, 1.39)
62%	140	3.2	0.843	(0.60, 1.18)
81%	183	5.4	0.702	(0.52, 0.94)
96%	218	10[illegible]	0.651	(0.50, 0.85)
100%	228	22[illegible]	0.664	(0.51, 0.87)

Overall follow-up is low even with 80% events

Treatment effect emerges late in the trial

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: in presence of NPH

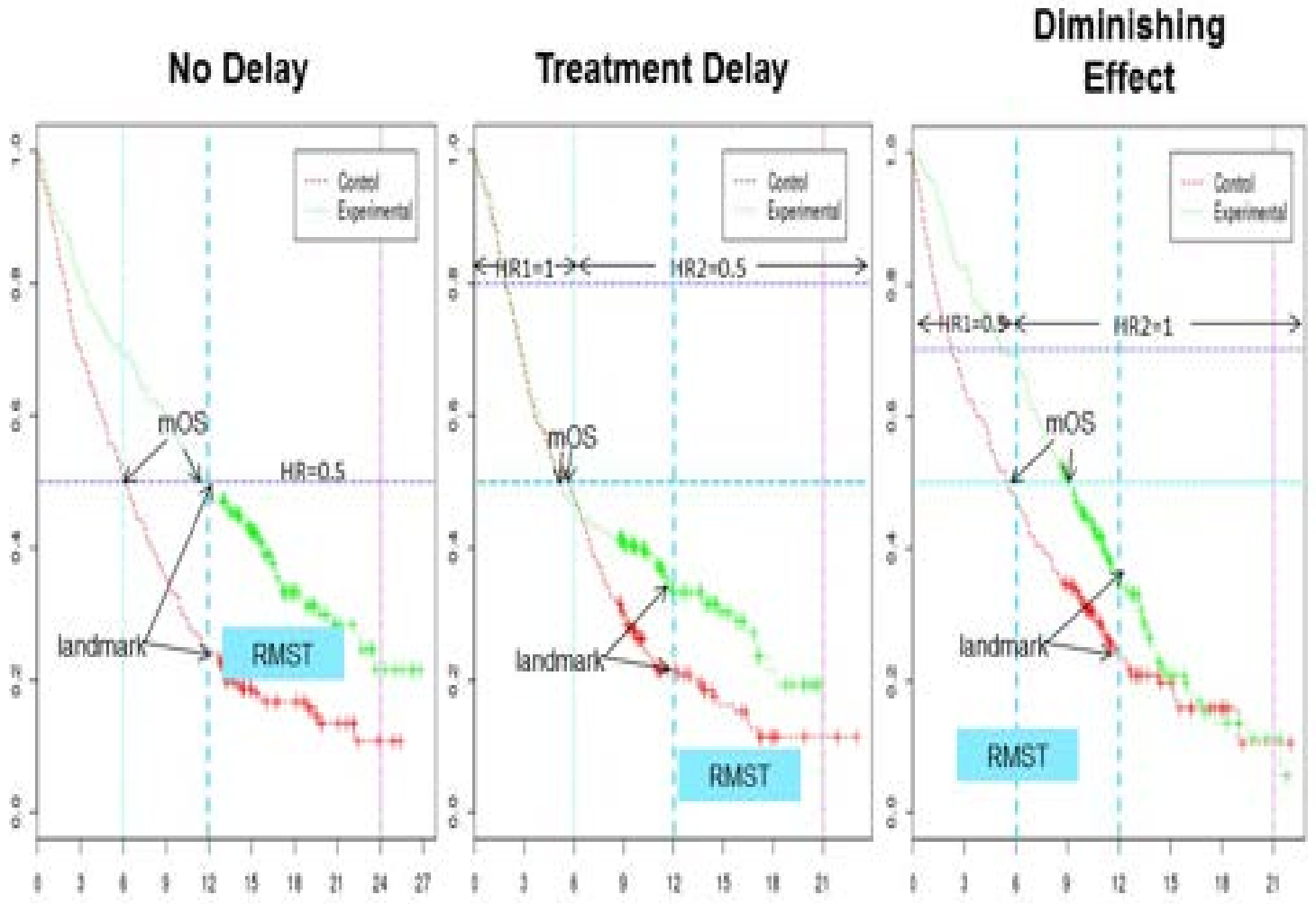
- How to plan primary analysis appropriately?
- How to design a trial?
- How to efficiently communicate the results with non-statisticians?

Hypothesis Testing: Best Way to Eliminate “Null” Effect

- Log-rank test is nonparametric (“rank based”) in nature => no assumptions related to shape of survival function or treatment effect
- Most powerful for detecting the alternatives with constant treatment effect
- May suffer significant power loss if treatment effect is not constant
- Other alternatives
  - How to tackle different potential alternatives?
  - Ensuring type-I error control

Estimation: Best Way to Describe the Results

- Hazard ratio
- Median
- HR over time
- Restricted Means
- Survival Time (RMST)
- Milestone



Important Issue to Address

Test of hypothesis	Is log-rank test adequate? Best method to use in presence of NPH?
Estimation	Is HR still the best way to describe the results? If not, then what other information is needed?
Trial design	How to size and power of study adequately? How to plan interim analyses?
Communication	How to communicate with non-statisticians

Analysis of Trial with NPH

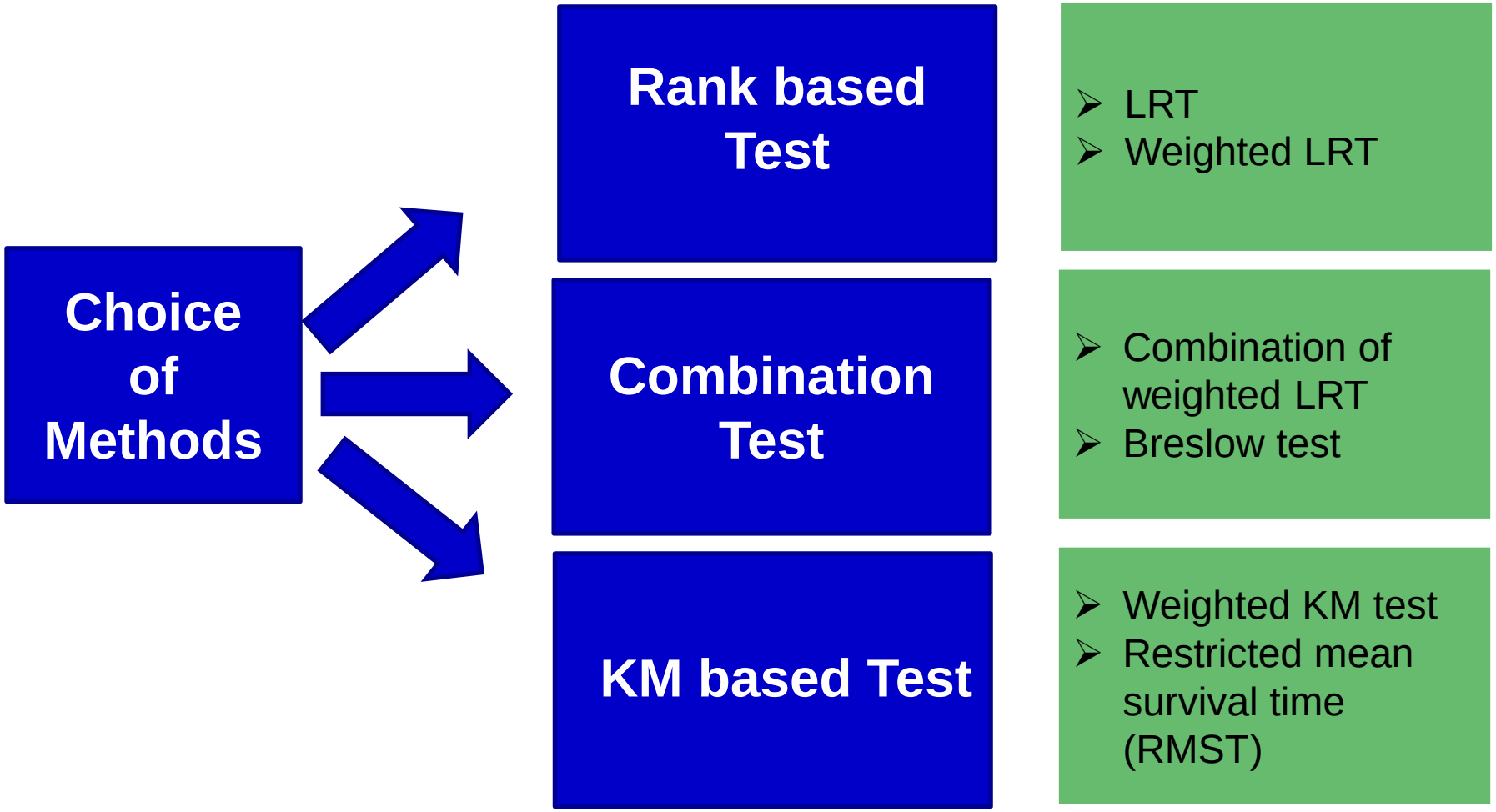
Choice of Primary Analysis in Confirmatory Trials

- 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)

Choice of Primary Testing Methods



Log-rank and Weighted Log-rank Test

- Log-rank test assumes that every point in time has the same relevance
  - Questionable under NPH
- Weighted Log-rank (WLR) attach a weight w_j with each points
- `Hard to specify w_j when NPH type is unknown

$$WLR = \frac{\sum_{j=1}^{D} w_j(O_j - E_j)}{\sqrt{\sum_{j=1}^{D} w_j^2 V_j}} = \frac{U(w)}{se(U(w))}$$

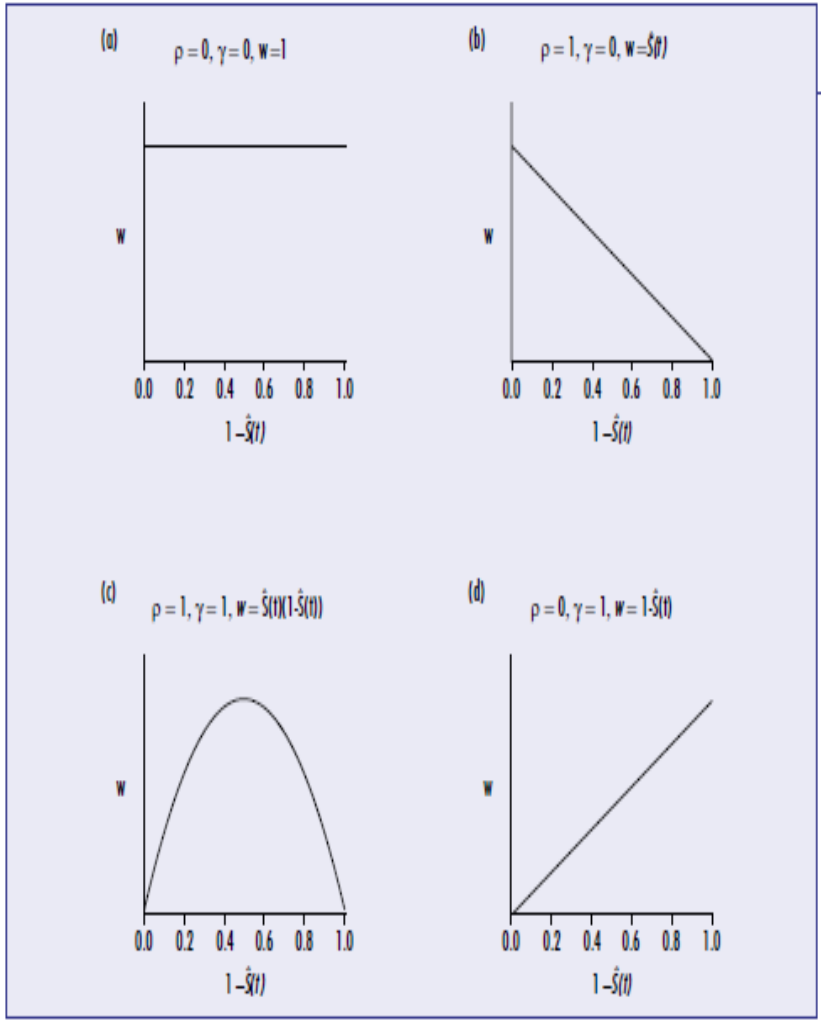
$$U(w) = \int_0^T w(t)\left(dN_1(t) - Y_1(t)\frac{dN(t)}{Y(t)}\right)$$

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

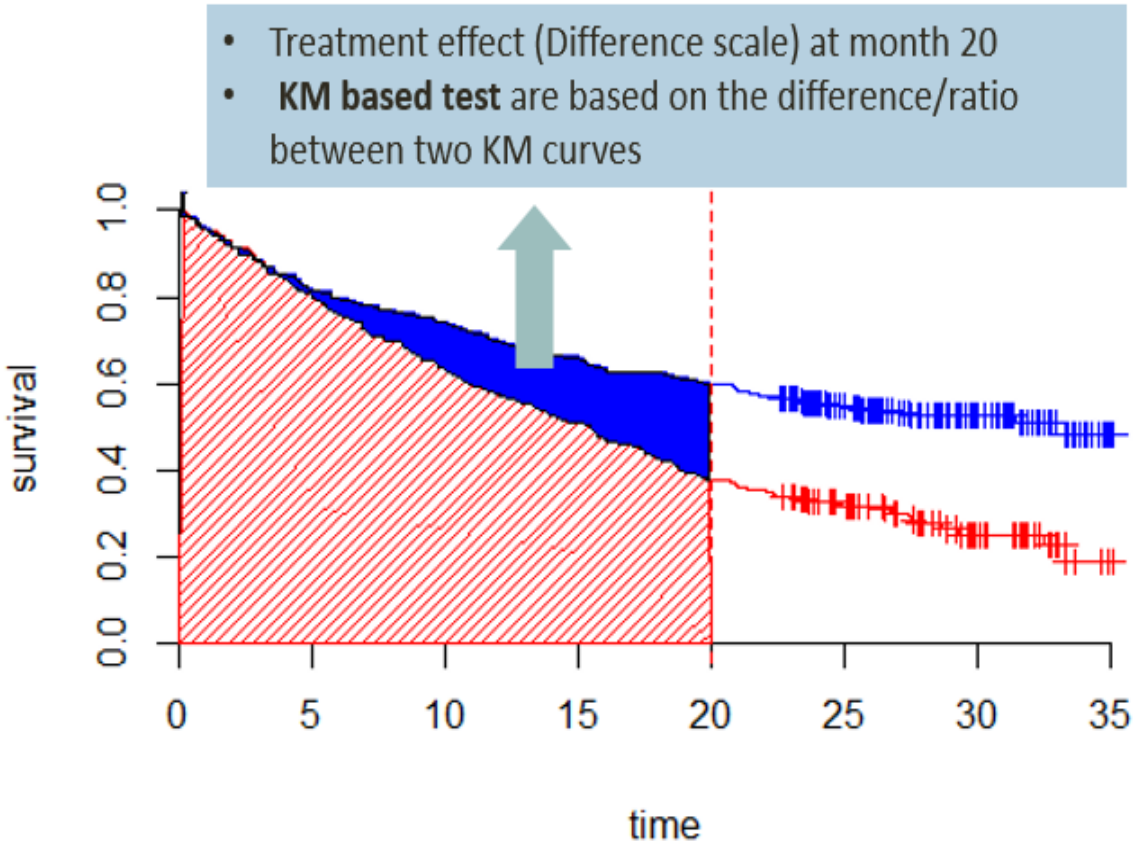


Other Methods

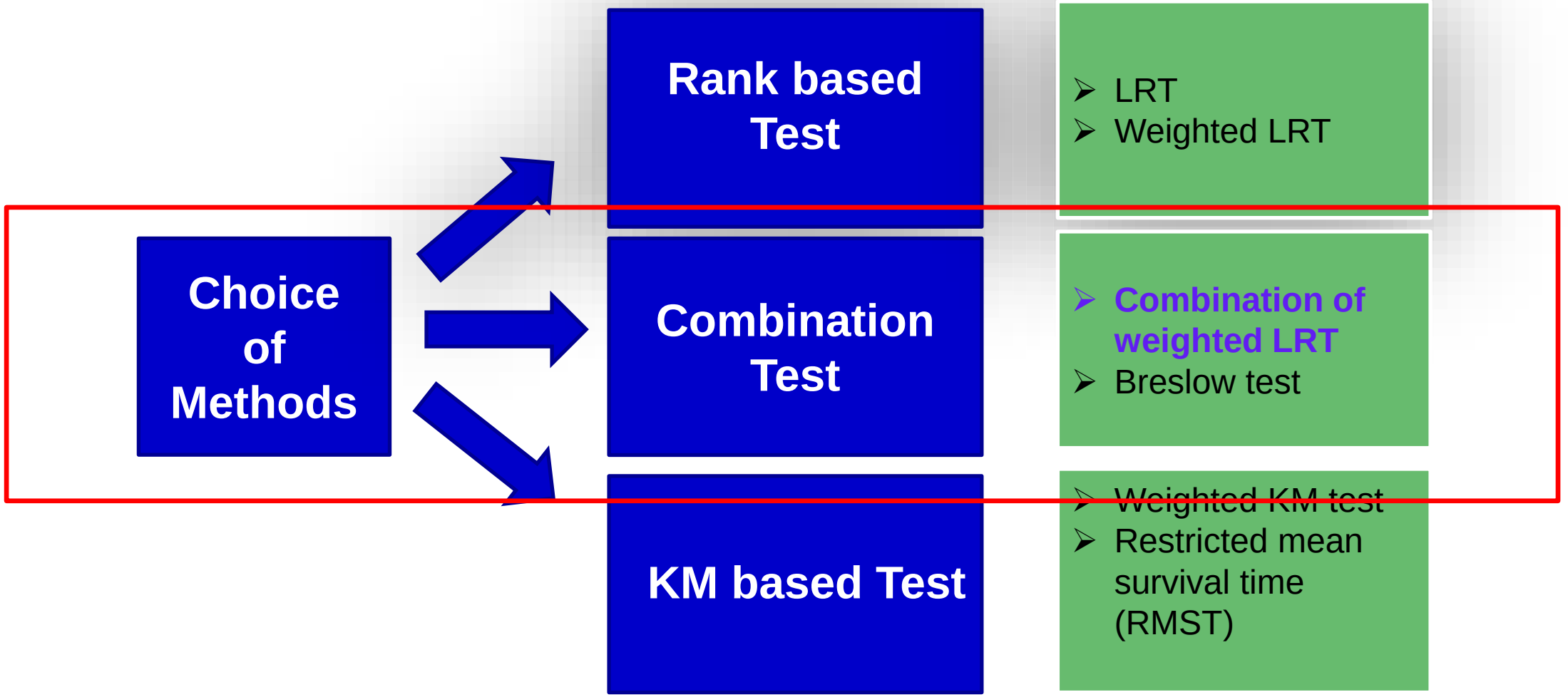
- Cox Regression with Time Dependent Coefficient (CoxTD): Putter 2005
- Piecewise LR Test (pWLRT): Xu et al. 2017, 2018
- Modestly Weighted LR Test (mWLRT): Magirr and Burman 2019

Kaplan-Meier (KM) Based Tests

- Based on the weighted KM statistic of two groups
  - WKM test (Pepe and Fleming 1983)
  - Restricted mean survival time (RMST)
  - Milestone survival at pre-defined timepoint
- Dependent of the follow-up time τ
- Performance of RMST depends on censoring pattern and choice τ
  - Data-dependent: unknown at the design state



Combination Test: Handling Wide Number of Alternatives



Combination Test

Handle a broad class of alternative hypothesis: Lee (2007), Karrison and others (2016), Breslow, Edler, and Berger (1984)

Considers multiple test statistics: choose best test statistics based on data

- Breslow, Edler, and Berger (1984): combination of LR test and test of acceleration
- Logan, Klein, and Zhang (2008): combination of LR test and milestone survival
- Lee (2007): Average and maximum of LR test (FH(0, 0)) and FH(0, 1)

Requires appropriate multiplicity control due to the correlation of test statistics

Often provides robust power under wide class of alternative hypotheses

Robust MaxCombo Test: A Potential Alternatives

- Proposed by Cross-Industry WG (Roychoudhury et al. (2020), Lin et al. (2020))
  - Motivated from the work from Yang and Prentice (2010) and Lee (2007)
- Based on multiple FH-WLR test statistics and chooses the best one adaptively depending on the underlying data
- We had proposed the following combination tests: largest of the test statistics
  - **Original MaxCombo test**: $G^{0,0}$, $G^{0,1}$, $G^{1,0}$, $G^{1,1}$
  - **Modified MaxCombo test**
    - **Option 1**: $G^{0,0}$, $G^{0,0.5}$, $G^{0.5,0}$, $G^{0.5,0.5}$: conservative and less sensitive to tail events
    - **Option 2 :** $G^{0,0}$, $G^{0,0.5}$, $G^{0.5,0.5}$: If delayed effect is only possibility

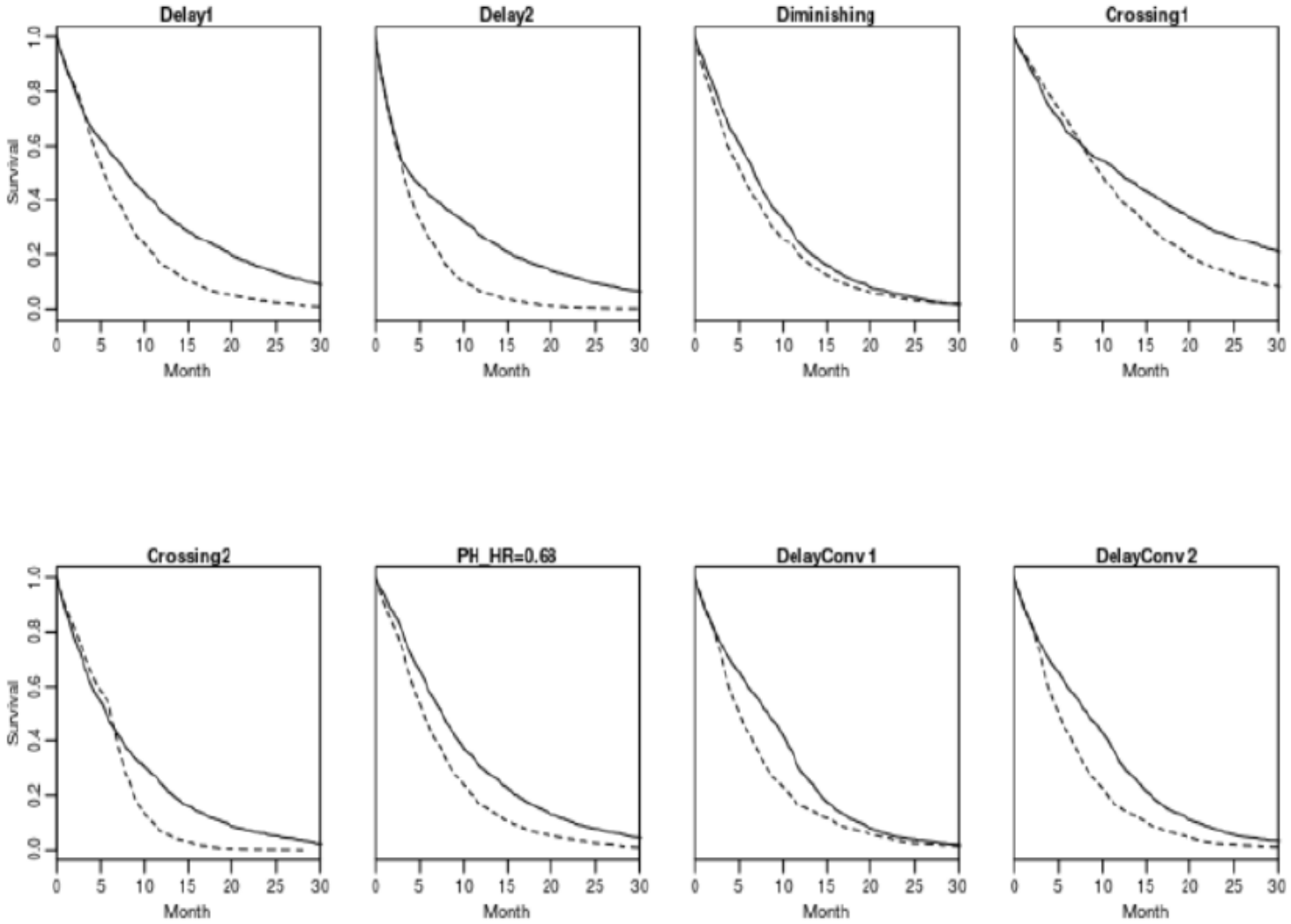
Practical Implementation

- “Adaptive” procedure involving selection of best test statistics: **requires multiplicity correction**
- Adjustment using the joint asymptotic distribution of the FH Log-rank test statistics
  - o Correlation between FH Log rank test statistics are analytically tractable
  - o Karrison et. al. (2016) proved asymptotic normality of the joint distribution
- MaxCombo p-value can be calculated using multivariate normal calculation
  - o Calculation can be done easily using R or SAS (R package: nphsim, simtrial)

Performance of MaxCombo Test: A Comparative Study

Rank based Test	➢ LRT ➢ FH Weighted LRT
Combination Test	➢ MaxCombo ➢ Breslow test
KM based Test	➢ Weighted KM test (WKM) ➢ Restricted mean survival time (RMST)

Performance of MaxCombo Test: Simulation Set-up



.

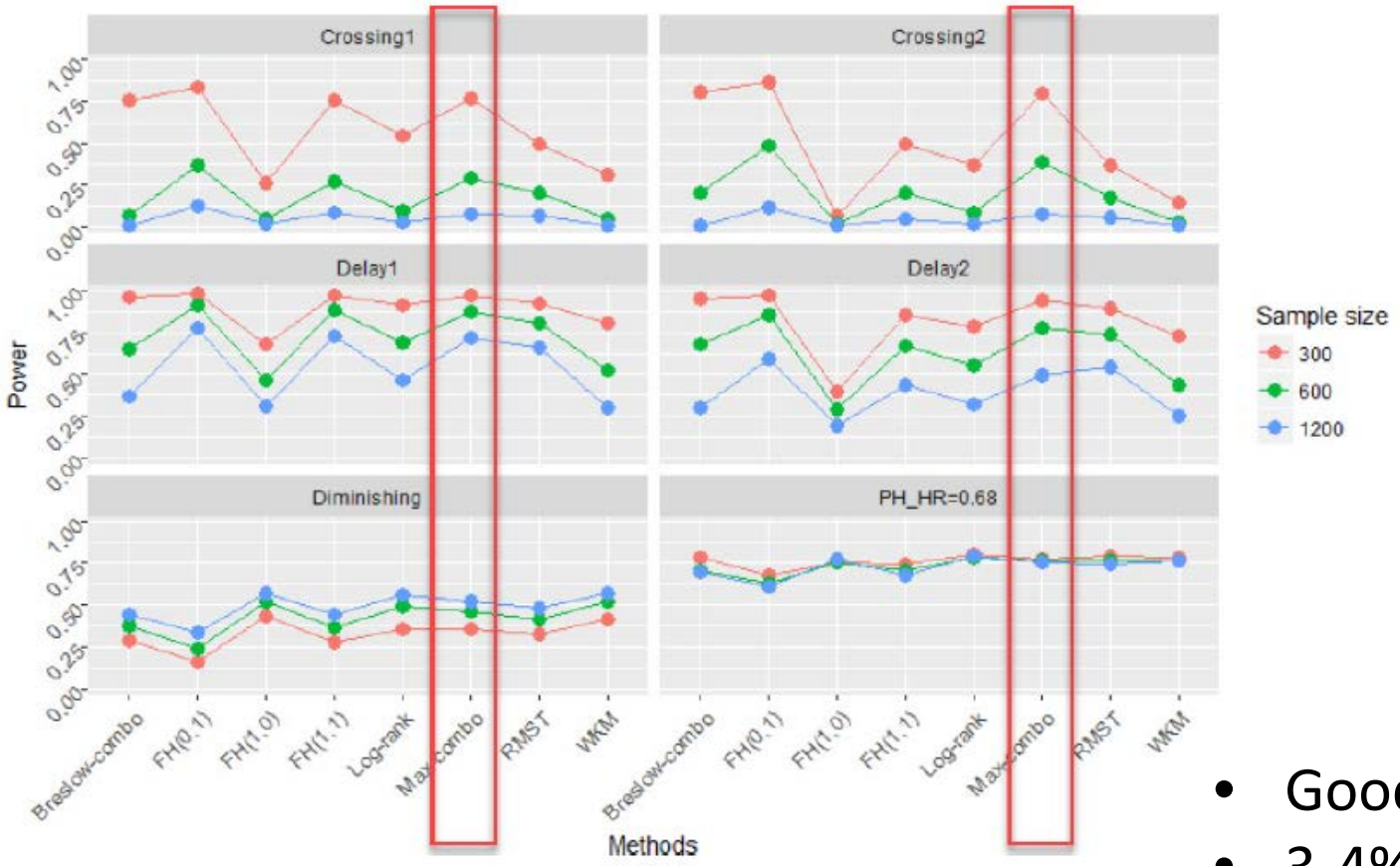
- **Dropout hazard rate**: λ=0.014
- **Sample size:** 300, 600, 1200
- **Enrollment:** 12, 18, 24 months
- **Number of event:** 210

Type 1 Error

- 20,000 trial datasets are simulated for each scenario under $H_0 : S_1(t) = S_0(t)$
- Type-I error is well protected with MaxCombo test with null

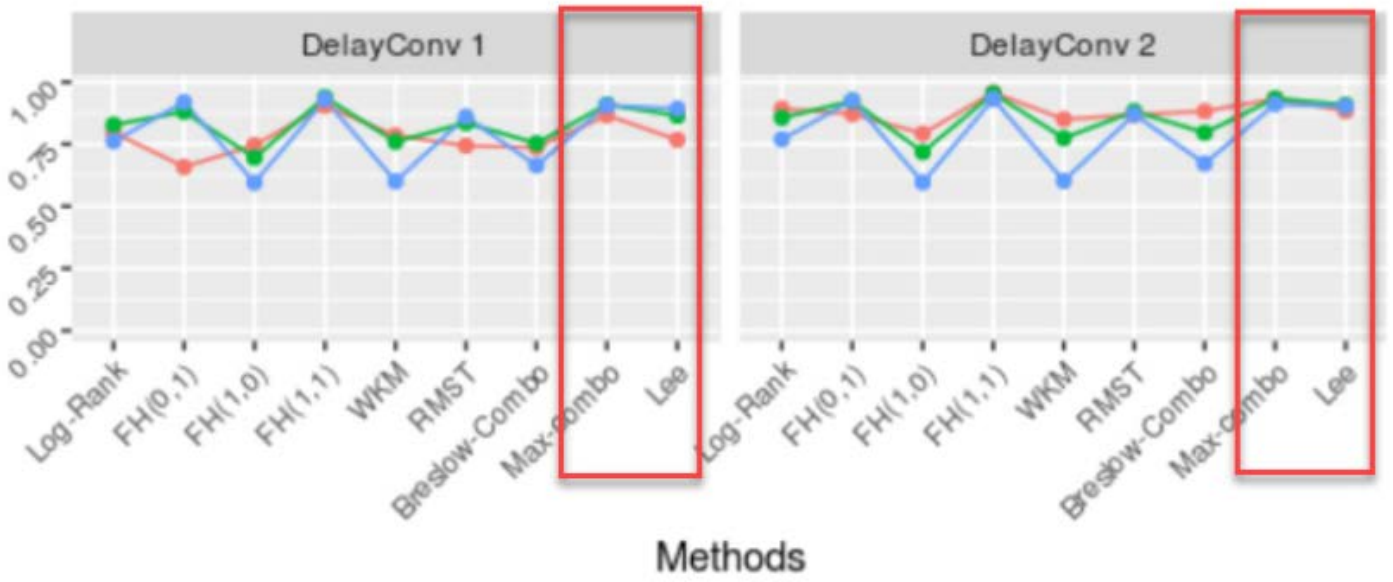
Sample Size	Log.Rank	FH(0,1)	FH(1,0)	FH(1,1)	RMST	WKM	Combo.Breslow	MaxCombo
300	2.590	2.630	2.520	2.605	2.545	2.575	2.505	2.595
600	2.585	2.430	2.770	2.380	2.590	2.730	1.210	2.415
1200	2.495	2.450	2.605	2.485	2.565	2.635	1.325	2.590

Empirical Power under Different NPH Setting



- Good power across different NPH scenarios
- 3-4% power loss under PH scenario

Advantage Over Existing Combination Test

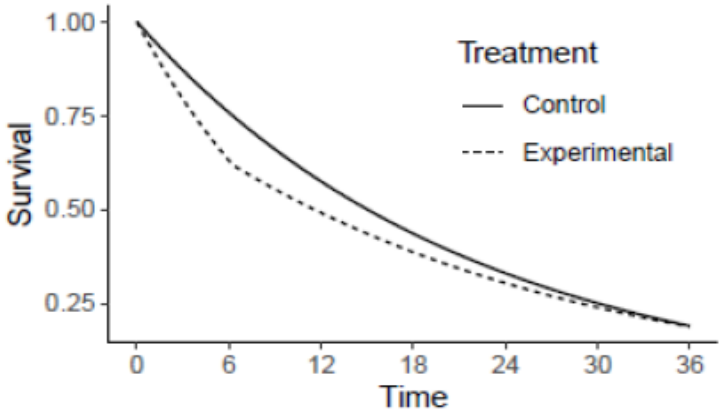


MaxCombo test has improved power over Lee test under delayed effect with converging tails scenarios

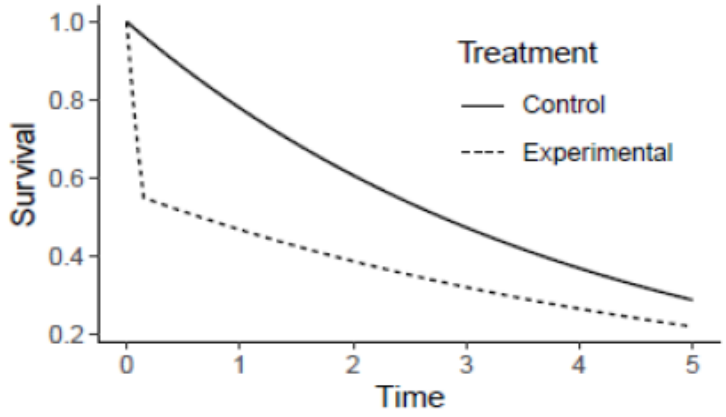
MaxCombo Test and Strong Null Hypothesis

- There are some concerns regarding the performance of the MaxCombo test under the *strong null* and *severe late crossing scenarios*
- Possibility of high probability of rejecting null hypothesis when the experimental drug is harmful
- Few additional scenarios are considered
  - *Strong Null 1* : Magirr and Burman (2019)
  - *Strong Null 2* : Freidlin and Korn (2019)
  - *Severe late crossing*: Treatment shows a late and marginal survival benefit
- The final cut-off date for each simulation is the calendar time of 5 years
- Should not be mixed with formal type-I error assessment

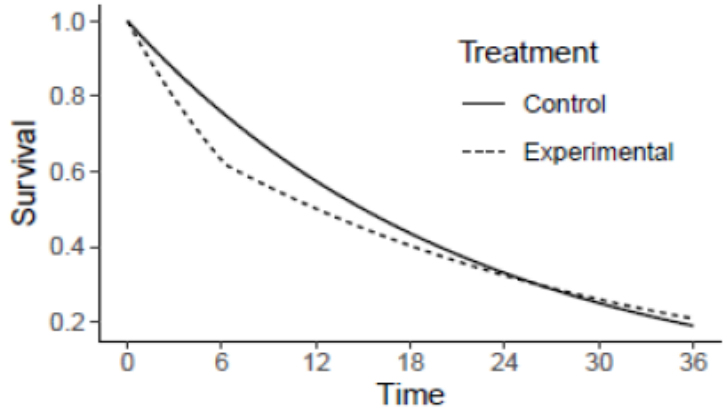
MaxCombo Under Extreme Scenarios



(a) Strong null 1



(b) Strong null 2



(c) Severe late Crossing

MaxCombo: Probability of Rejecting Null Hypothesis Under Extreme Scenarios

Strong Null 1

- Recruitment uniformly over 12 months: **2.1%**
- Recruitment uniformly over 6 months: **2.3%**

Severe Crossing

- Recruitment uniformly over 12 months: **5.0%**
- Recruitment uniformly over 6 months: **5.8%**

Strong Null 2

- Recruitment uniformly over 12 months: **39.0%**
- Recruitment uniformly over 6 months: **48.9%**

Strong Null 2 and Modified MaxCombo Test ($G^{0,0}$, $G^{0,0.5}$, $G^{0.5,0}$, $G^{0.5,0.5}$)

Modified MaxCombo can handle scenario like Strong Null 2

- **Strong Null 1**: **0.05%**
- **Strong Null 2**: **1.8%**
- **Severe Crossing**: **2.6%**

Such scenarios are unrealistic in real-life: will be stopped early by a data monitoring committee (DMC) due to the safety concerns

Additional simulations showed robust power for the modified MaxCombo

- Showed better empirical power than Log-rank and mWLRT tests under delayed effect and crossing survival (Roychoudhury et al. (2020))

Primary Analysis for Confirmatory Trials

- Under NPH, no single efficacy measure is sufficient
- A p-value from any single statistical test or a single summary statistic fails to capture treatment benefit
- A robust testing procedure like MaxCombo or modified MaxCombo test is required to handle uncertainties associated with NPH type
- Additional pre-specified measures beyond HR and median needed to describe benefit over entire follow-up period; e.g., milestone survival, RMST
- Ensure adequate follow-up to evaluate time-dependent treatment effect

Stepwise Approach for Primary Analysis

A stepwise approach for primary analysis in trials where NPH is expected

- **Step 1:** Perform a statistical test to reject “Null” hypothesis (no treatment effect)
- **Step 2**: Evaluate PH assumption using standard methods
- **Step 3:** Select treatment effect summary based on step 2 findings
  - if PH is reasonable: use traditional measures like HR and median
  - if PH is not reasonable: also provide additional measures such as **milestone survival rate, restricted mean survival time (RMST)** (piecewise HR at pre-specified time points as exploratory)

This approach provides a complete summary of any treatment effect

Appropriately pre-specification in protocol and SAP is possible to meet ICH E9

Modified MaxCombo as Primary Test

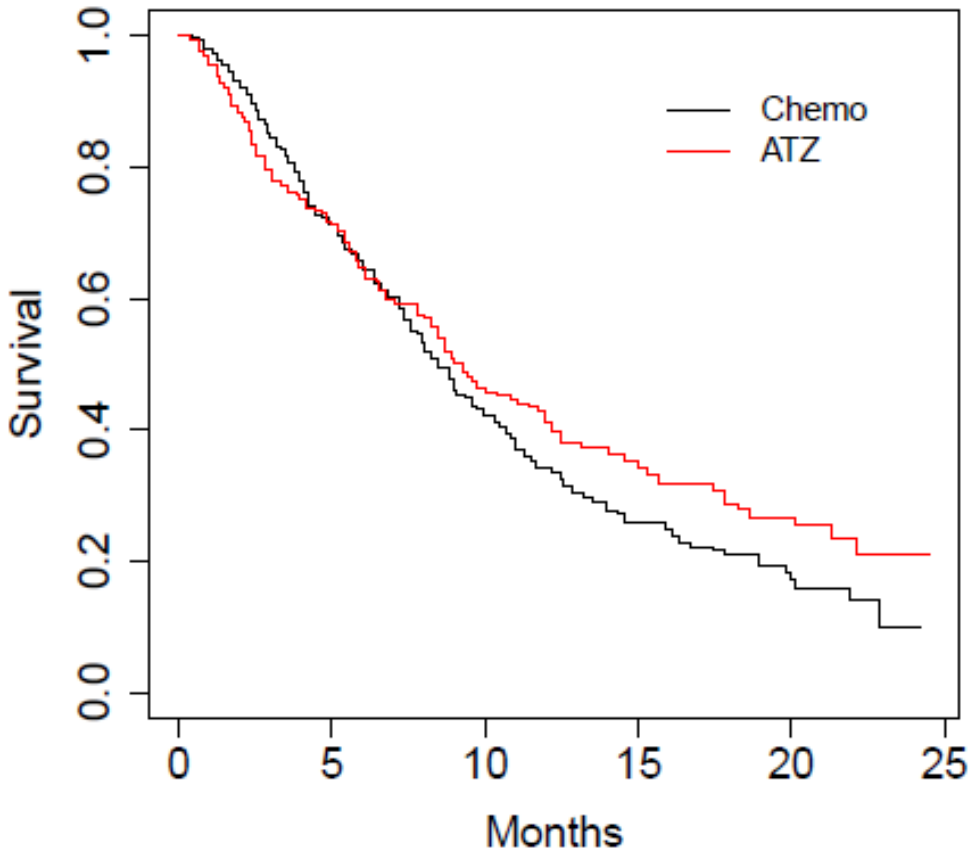
Use **modified MaxCombo** for primary statistical testing: A combination test based on Fleming-Harrington weighted LRT

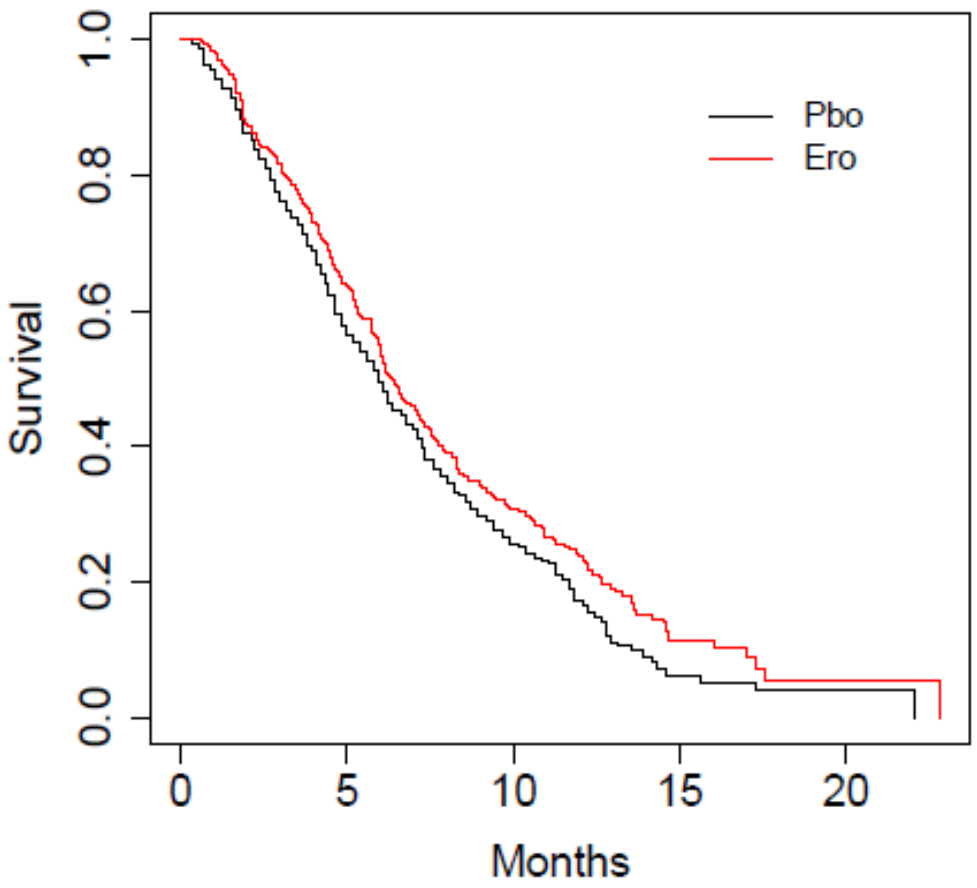
- Considerable loss in power of LRT under NPH
- Extensive simulation study shows better statistical power of the modified MaxCombo test over traditional LRT under various types of NPH (especially for delayed treatment effect)
- Maintains good statistical properties under PH
- Can be pre-specified in SAP

LRT still recommended to be used as a supportive evidence

Under potential NPH, design should specify sample size and total follow-up time to ensure adequate power

Example: Overall Survival IM211 Trial IC1/2/3 Cohort (Digitized) and PA3 Trial (Digitized)





Application of Stepwise Approach for IM211 and PA3 Overall Survival

Method	IM211: Digitized	PA3: Digitized
Traditional Analysis		
LR test	0.040	0.023
HR and 95% CI	0.847 (0.70, 1.02)	0.834 (0.70, 0.99)
Median (month)	8.9 vs 8.3	6.2 vs 5.9
Stepwise Approach		
MaxCombo	0.005 (FH(0,1))	0.048 (FH(0,0))
WHR and 95% CI	0.731 (0.57, 0.93)	0.834 (0.68, 1.03)
Difference in RMST and 95% CI	1.090 (-0.22, 2.40)	0.860 (-0.07, 1.79)
Difference in milestone rates at 12 months	0.021 (-0.04, 0.18)	0.083 (-0.01, 0.15)

Design of Trial with NPH

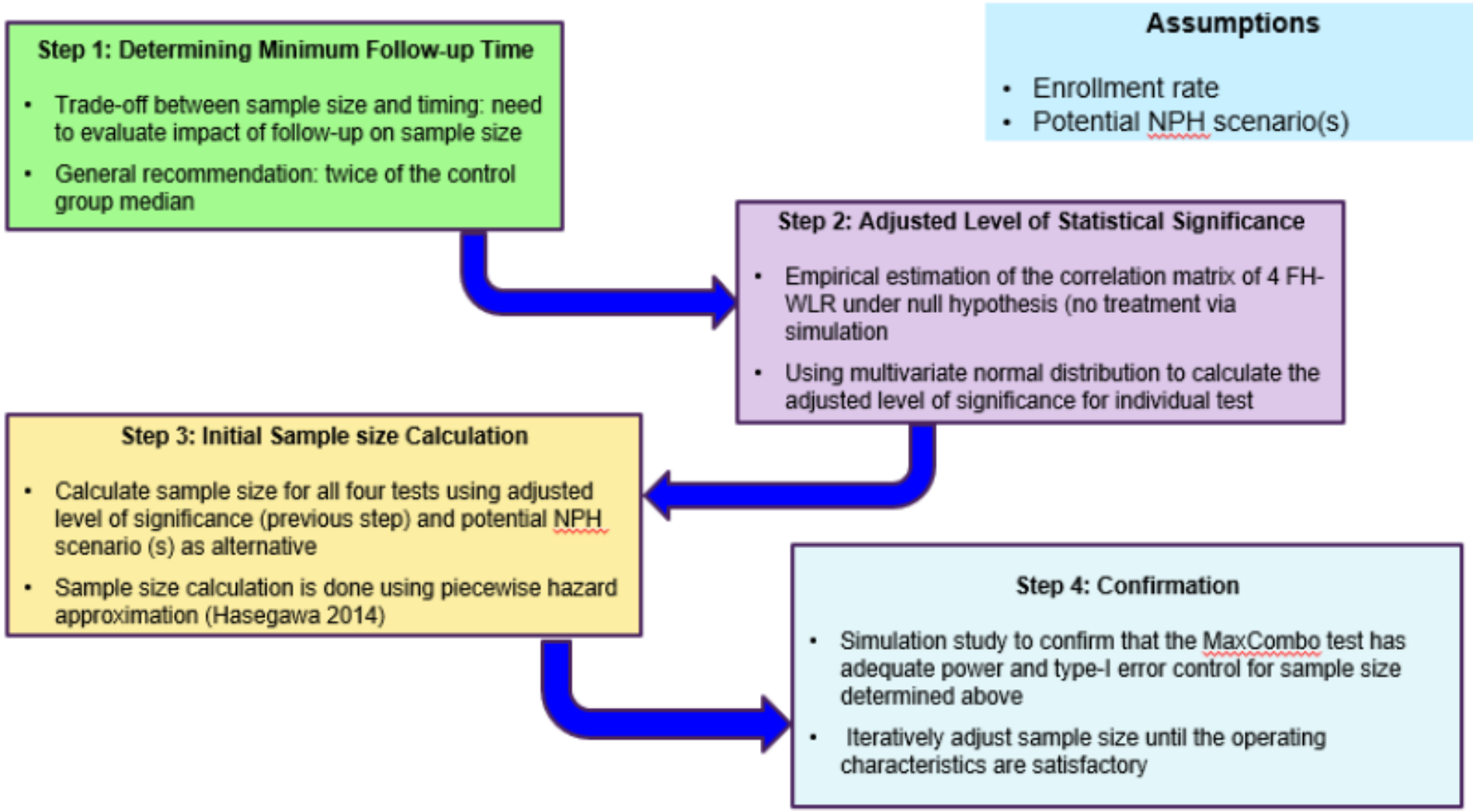
Design Challenges with Potential NPH

- Potential of NPH brings lot more uncertainties in design assumption
- Treatment differences under NPH constitute a broad class of alternative hypotheses
  - Degree of effect
  - Delayed timing of effect: Delayed separation of survival curves
  - Different effects in unanticipated subpopulations: Can result in crossing hazards
  - Diminishing effect over time
- How do we design a trial to be powerful across MANY alternatives?
- We focus on trials with high early event rate (e.g. metastatic cancer)

Design with Potential NPH : General Considerations

- Trial duration or total follow up time plays an important role
  - Event based only analysis may produce a design that finishes too early
  - Underpowered
  - May fail to describe time dependent treatment effect
- Carefully elicitation of the possible treatment effect scenario
  - Power trial for multiple scenarios
    - Protocol/Analysis plans should present at least 2 relevant scenarios for which a trial is well powered
  - Plan for worst-case scenarios
    - Delayed treatment effect for IO trials

Sample Size Calculation: Two Step Approach



Group Sequential Design with MaxCombo Test

- Use of log-rank test for interim analysis and MaxCombo for final analysis
  - To avoid the impact of short follow up time or trial duration in WLR
  - Well accepted by the regulators
- Final success boundary needs multiplicity adjustment due to the correlation between the LR test at interim and the MaxCombo test in final analysis
- Calculation of the final boundary using independent increment of information from interim to final and asymptotic normality
- The impact on type-I error and power for interim analysis need to be evaluated via simulation

Cross Pharma NPH Working Group

- A cross industry collaboration group is formed in collaboration with FDA
  - FDA division of Biometrics V (Oncology) and IX (Hematology) are closely involved
- Achievements and timelines
  - Initial kick off meeting**:** ASA BIOP RISW, October 2016
  - Work stream formed: January 2017
  - Face to Face midpoint meeting: ASCO 2017
  - Duke-Margolis public workshop: February 2018
  - FDA meeting to communicate key findings and recommendations: November 2019
  - Two papers published, Two in progress
  - Multiple presentation and training in major statistical conference
  - Two R packages: nphsim and simtrial (available in GitHub)

Summary

MaxCombo 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%)
- Additional caution is required while interpreting the severe crossing cases

A three-step primary analysis to provide summary of appropriate treatment effect

Design with MaxCombo test requires both sufficient events and follow-up

- Asymptotic methods and simulation are currently required for planning

Early analysis is problematic when treatment effect changes over time

- We propose a group sequential strategy based on log-rank and MaxCombo which is intuitive and regulatory appealing

NPH Group References

- Roychoudhury, S., Anderson, K.M., Ye, J & Mukhopadhyay, P. (2021). Robust Design and Analysis of Clinical with Trial Non-proportional Hazards: A Straw Man Guidance. **Statistics in Biopharmaceutical Research** https://doi.org/10.1080/19466315.2021.1874507

- Lin R, Lin J, Roychoudhury S et al. (2020) Cross-pharma non-proportional hazards working group. alternative hypothesis analysis methods for time to event endpoints under non-proportional hazard: A comparative analysis. **Statistics in Biopharmaceutical Research** tps://doi.org/10.1080/19466315.2019.1697738

- R Packages
  - o simtrial: https://github.com/keaven/simtrial
  - o gsDesign2: https://rdrr.io/github/keaven/gsDesign2/
  - o gsdmvn: https://github.com/keaven/gsdmvn/

Thank You