Impact of Non-proportional Hazard in Confirmatory Trial Design and Analysis for Oncology

Satrajit Roychoudhury

October 24, 2017

Pfizer GLOBAL PRODUCT DEVELOPMENT

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

- Most popular methods for analysis of time to event trials:
  - log-rank test (unstratified and stratified)
  - Cox regression
  - Kaplan-Meier plot
- Hazard ratio and naive median are standard way of summarizing treatment effect
- Are they good summary measures when the treatment effect is not constant over time? : **Non-proportional Hazard (NPH) problem**
  - For example, recent immunotherapy development showed evidence of a delayed effect
- How to cope with NPH problem at design and analysis stages?

Cox Proportional Hazard Model: Gold Standard

- Introduced by Sir David R Cox in 1976
- Model contains two parts
  - baseline hazard : risk of event per time unit changes over time
  - covariate effect : constant over time -> **Proportional Hazard (PH)**
  - two components are multiplicatively related
  - treatment effect is typically communicated via hazard ratio (HR)
- D. R. Cox also introduced partial likelihood method: covariate effects can be estimated without estimating baseline hazard
- However, the biological interpretation of PH is often tricky
- Failure to meet PH assumption causes the interpretation of the study result challenging

Log-rank Test: A Remedy to All?

- Introduced by Nathan Mantel and named by Richard and Julian Peto
- Log-rank (LR) doesn't assume PH as makes less 'assumptions' as it's fully nonparametric
- However: LR test and Cox model are closely related
  - LR test provides identical results to score test from the cox model, if the methods to handle ties and stratification are same
- Power of log-rank is most powerful for proportional hazards type alternatives

$$H_1 : S_1(t) = S_0(t)^{\exp(\beta)} \Leftrightarrow h_1(t) = h_0(t)e^{\beta}$$

- Can cause substantial power loss if PH assumption does not hold

NPH: What is Really New ?

- The topic NPH has been discussed extensively in statistical literature
- There are number of alternatives and extensions are proposed
  - weighted log-rank test, time-dependent cox regression, accelerated failure time model, restricted failure time model and many others ….
- Use of NPH methodologies are rare in real life clinical trial
  - ~98% of the clinical trials with time to event endpoint uses log-rank test and cox PH model for design primary analysis (source: NEJM 2000-2017)
- How to cope with it?
  - how to perform design and analysis in regulatory environment?
  - how to efficiently communicate the results with nonstatisticians?

NPH in Oncology Trials: What Brings It to Forefront?

- Immuno-Oncology : rapidly evolving field
- IO agents can impact the immune system as well as the tumor microenvironment
  - tumors may be eradicated from the host or delay disease progression
- NPH is not new in Oncology
  - similar incidence has been observed in adjuvant and neoadjuvant trials
- FDA Guidance on *Clinical considerations for therapeutic cancer vaccines* in 2011 states:

  "this delay in the effect may lead to an average effect that is smaller than expected and thus may require both an increase in sample size to compensate for the delay and a careful assessment of trial maturity for the primary analysis."

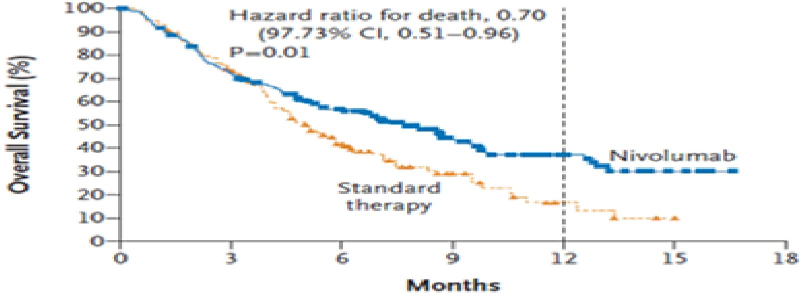
1 Delayed Treatment Effect

CM – 141 (2L SCCHN)

http://www.nejm.org/doi/full/10.1056/NEJMoa1602252#t=article

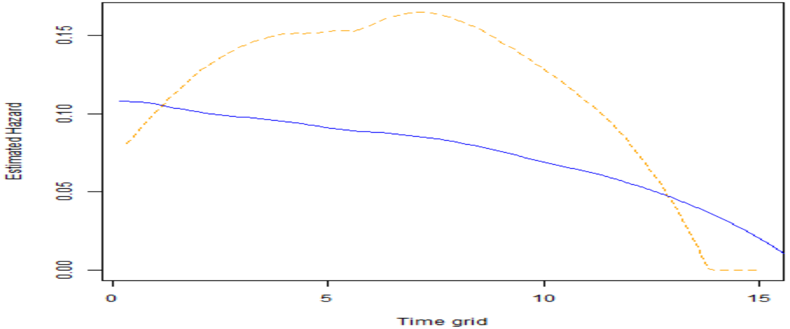
A Overall Survival

	No. of Patients	No. of Deaths	1–Yr Overall Survival Rate % (95% CI)	Median Overall Survival mo (95% CI)
Nivolumab	240	133	36.0 (28.5–43.4)	7.5 (5.5–9.1)
Standard Therapy	121	85	16.6 (8.6–26.8)	5.1 (4.0–6.0)



No. at Risk							
Nivolumab	240	167	109	52	24	7	0
Standard the	121	87	42	17	5	1	0

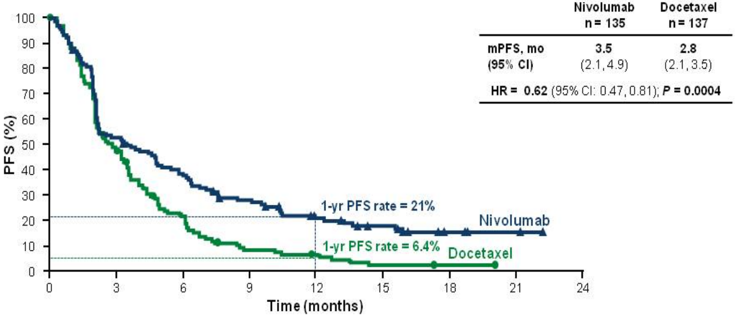
Checkmate141-OS



CM – 017 (2L Squamous NSCLC)

http://www.nejm.org/doi/full/10.1056/NEJMoa1504627#t=article

Progression-Free Survival



	Nivolumab n = 135	Docetaxel n = 137
mPFS, mo (95% CI)	3.5 (2.1, 4.9)	2.8 (2.1, 3.5)

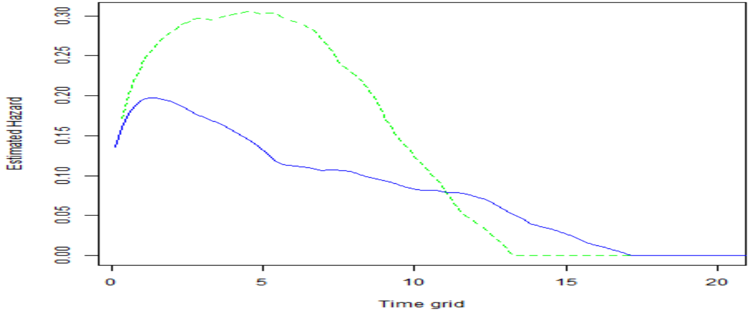
HR = 0.62 (95% CI: 0.47, 0.81); ***P*** **= 0.0004**

Number of Patients at Risk									
Nivolumab	135	68	48	33	21	15	6	2	0
Docetaxel	137	62	26	9	6	2	1	0	0

PFS per investigator.

PRESENTED AT: ASCO Annual '15 Meeting

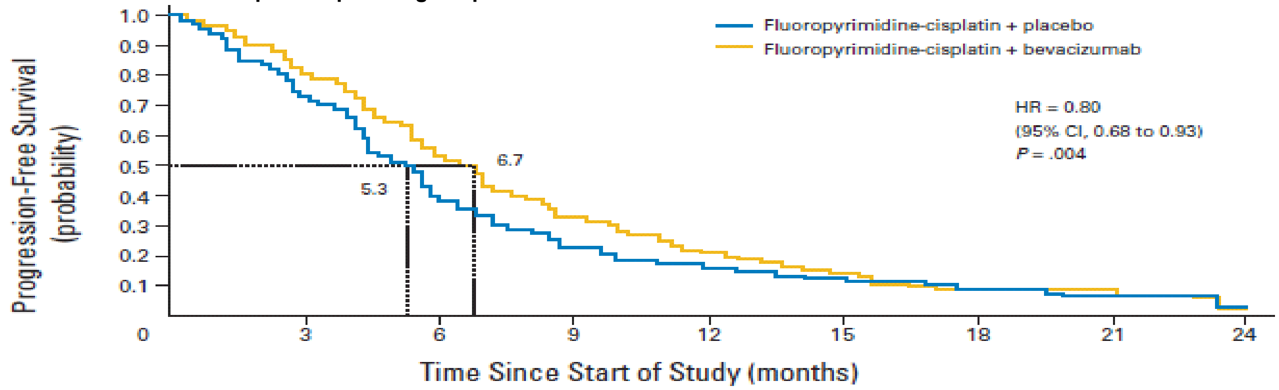
Checkmate017-PFS



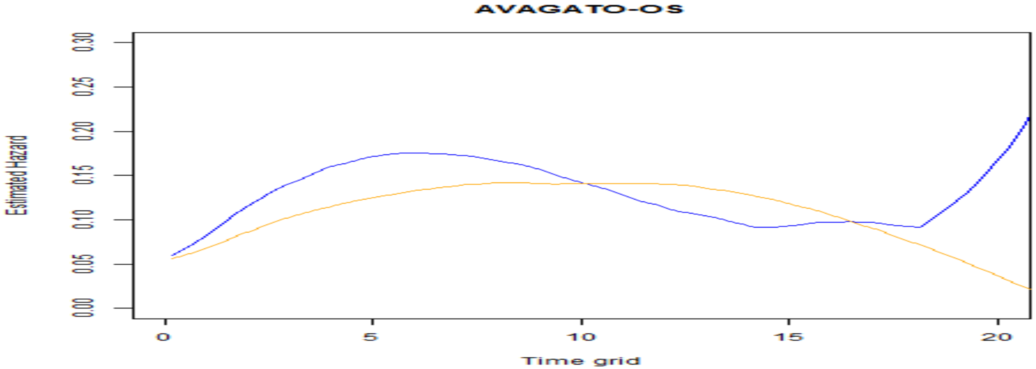
2 Diminishing Treatment Effect

AVAGAST Trial (Gastric Cancer)

http://ascopubs.org/doi/pdf/10.1200/JCO.2011.36.2236



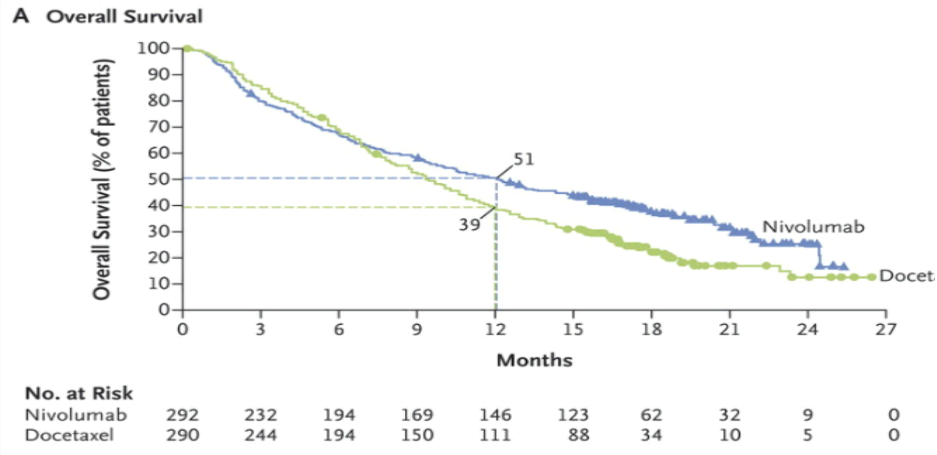
No. at risk									
Fluoropyrimidine-cisplatin + placebo	387	279	145	86	55	32	15	3	0
Fluoropyrimidine-cisplatin + bevacizumab	387	306	201	123	71	38	11	3	0



3 Crossing Hazard Functions

CM- 057 (Non-squamous NSCLC)

http://www.nejm.org/doi/full/10.1056/NEJMoa1507643#t=article

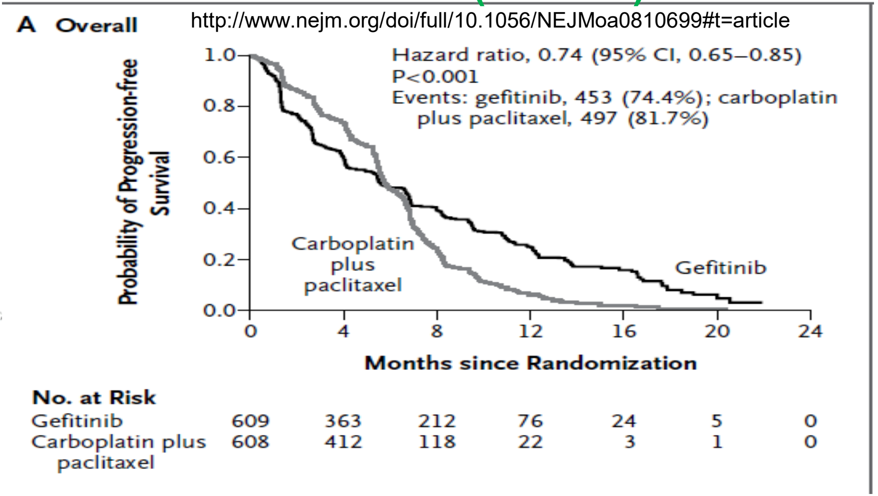


No. at Risk

Nivolumab	292	232	194	169	146	123	62	32	9	0
Docetaxel	290	244	194	150	111	88	34	10	5	0

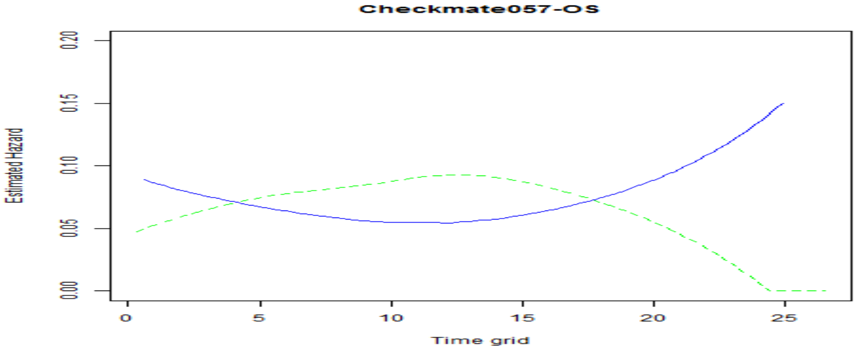
IPASS (NSCLC)

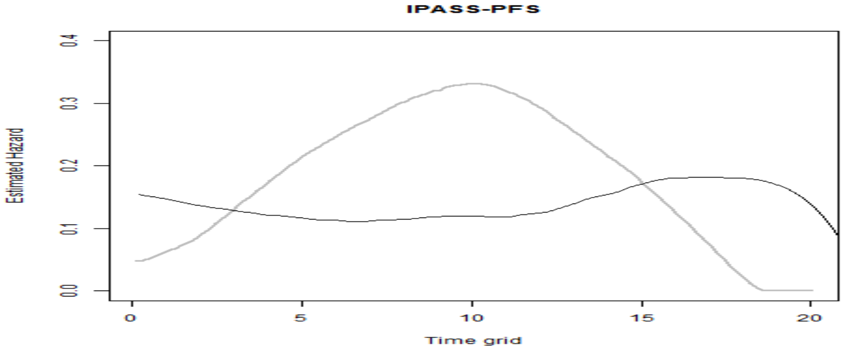
http://www.nejm.org/doi/full/10.1056/NEJMoa0810699#t=article



No. at Risk

Gefitinib	609	363	212	76	24	5	0
Carboplatin plus paclitaxel	608	412	118	22	3	1	0

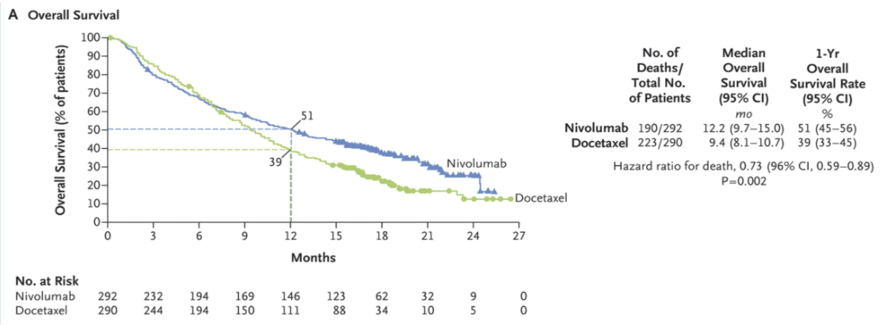




4 Biomarker Subgroups

CM- 057 (Non-squamous NSCLC)

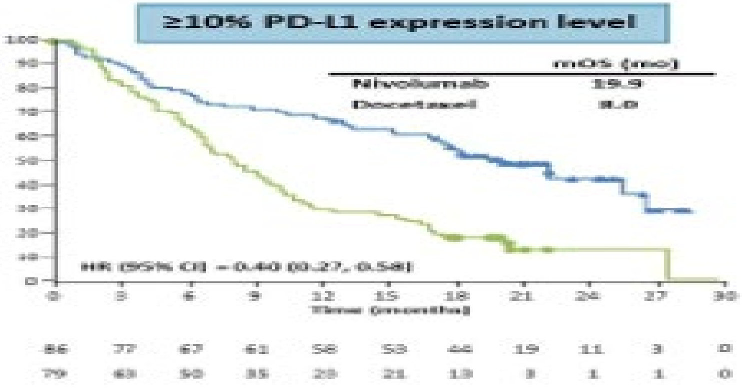
http://www.nejm.org/doi/full/10.1056/NEJMoa1507643#t=article



	No. of Deaths/ Total No. of Patients	Median Overall Survival (95% CI)	1-Yr Overall Survival Rate (95% CI)
		mo	%
Nivolumab	190/292	12.2 (9.7–15.0)	51 (45–56)
Docetaxel	223/290	9.4 (8.1–10.7)	39 (33–45)

Hazard ratio for death, 0.73 (96% CI, 0.59–0.89)
P=0.002

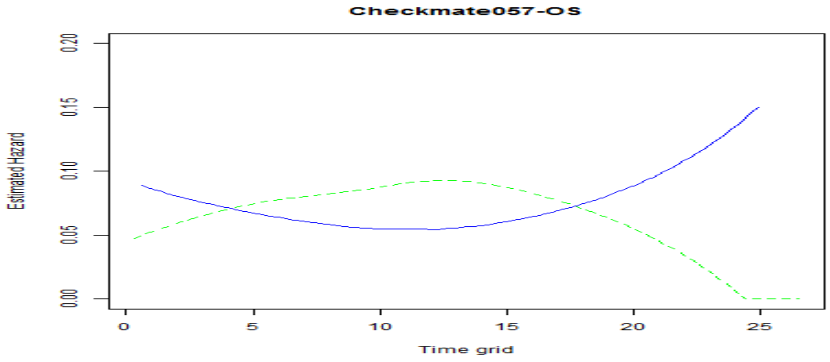
No. at Risk										
Nivolumab	292	232	194	169	146	123	62	32	9	0
Docetaxel	290	244	194	150	111	88	34	10	5	0

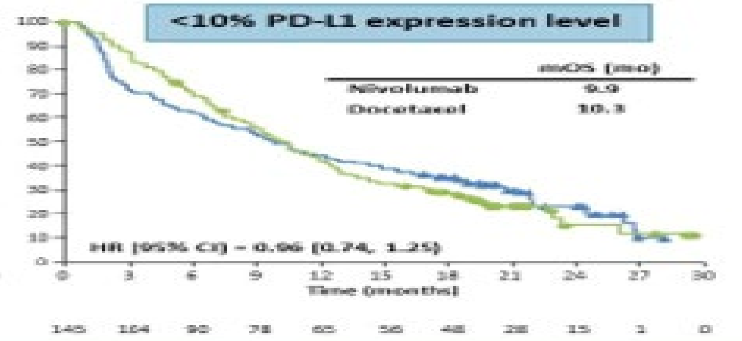


	mOS (mo)
Nivolumab	19.9
Docetaxel	8.0

HR (95% CI) = 0.40 (0.27, 0.58)

86	77	67	61	58	53	44	19	11	3	0
79	63	50	35	25	21	13	3	1	1	0





	mOS (mo)
Nivolumab	9.9
Docetaxel	10.3

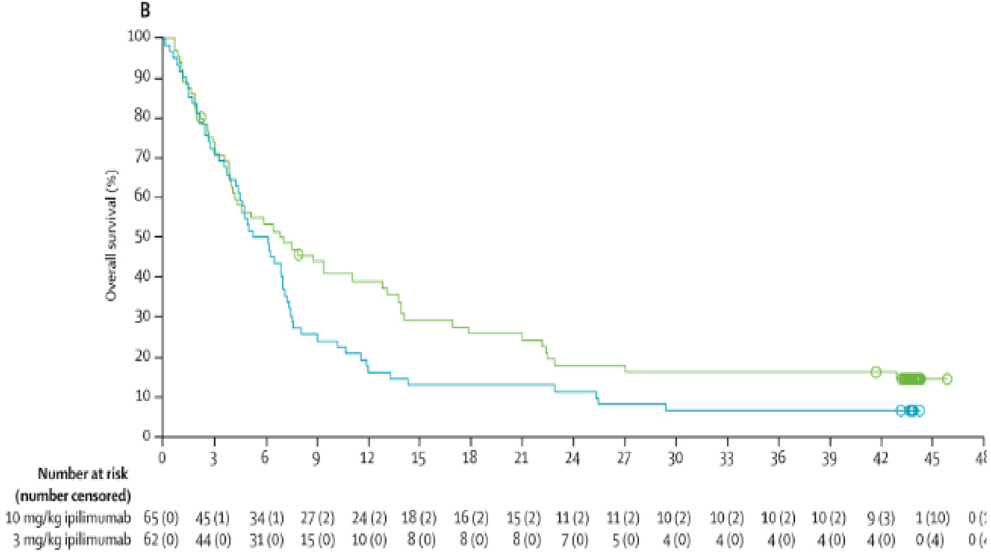
HR (95% CI) = 0.96 (0.74, 1.25)

145	104	90	78	65	56	48	28	15	1	0

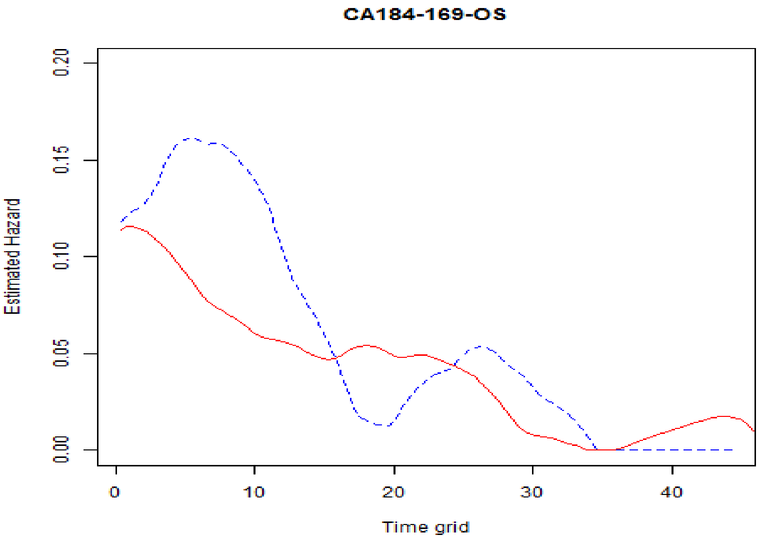
5 Delayed Treatment Effect + Long Term Survivor

CA184-169 (Metastatic Melanoma)

http://thelancet.com/journals/lanonc/article/PIIS1470-2045(17)30231-0/references



Number at risk (number censored)																	
10 mg/kg ipilimumab	65 (0)	45 (1)	34 (1)	27 (2)	24 (2)	18 (2)	16 (2)	15 (2)	11 (2)	11 (2)	10 (2)	10 (2)	10 (2)	10 (2)	9 (3)	1 (10)	0 (
3 mg/kg ipilimumab	62 (0)	44 (0)	31 (0)	15 (0)	10 (0)	8 (0)	8 (0)	8 (0)	7 (0)	5 (0)	4 (0)	4 (0)	4 (0)	4 (0)	4 (0)	0 (4)	0 (



Implementation in Practice: Points to Consider

- **Trial design**
  - size and power of study
  - planning follow-up
- **Analysis**
  - timing
  - interim analyses?
  - best method to use for testing and estimation in presence of NPH?
- **Interpretation**
  - is hazard ratio still the best way to describe the results?
  - if not, then what other information is needed?

Different Types of NPH

- Different types of NPH
  1. delayed treatment effect
  2. diminishing treatment effect
  3. crossing hazard
  4. subgroup effect: NPH driven by particular subgroup
  5. combination of 1/2/3 and long term survivor
- For today's presentation we'll concentrate on type 1 and confirmatory trial for illustration purposes
- Proposed methodology can be extended easily

Example: Standard Design

- Design set up: Treatment A vs SOC
  - progression Free Survival (PFS) as primary endpoint
  - median SOC: 5 months
  - alternative HR (Treatment A vs SOC): 0.67
  - enrollment period: 37.1 months
  - type I error 2.5%, power 90%
- Requires:
  - **297 (150 per arm)** patients
  - **263 events** to target : analysis timing ~ **44** months
  - minimum HR for statistical significance: **0.785**
  - planned primary analysis: log rank and cox regression

Effect of Delayed Treatment Effect (5000 simulation)

Delay in effect: 2 months

- Time dependent HR
  - HR = 1 t ≤ 2 months
    = 0.67 t > 2 months
- With **263 events**
  - analysis timing ~ 43 months
  - power : **~ 63% ↓**
  - HR estimate ↑
    - > 50% estimates ≥ 0.75

Delay in effect: 2 months

- With **520 events** and **600 patients**
  - analysis timing ~ 43 months
  - power: ~ 90%
  - the estimation problem remains
    - > 50% of HR estimates ≥ 0.75
- Significant increase in resources
  - Sample size **doubled**
- Cox regression does not seem appropriate

Choice of Primary for Confirmatory Trials

- 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 methods for
  - different type of NPH
  - different specifications (e.g. lag time for treatment effect)

Statistical Methods to Handle NPH

- Class of Log-rank and weighted Log-rank tests
  - log-rank test and weighted log-rank test
  - piecewise log-rank test (APPLE/SAPPLE)
  - supremum (Renyi)
- Class of test based on Kaplan Meier Curve
  - RMST
  - weighted KM Test
  - Milestone survival analysis
- Combination tests
- Parametric model

Log-rank and Other Rank Based Statistics

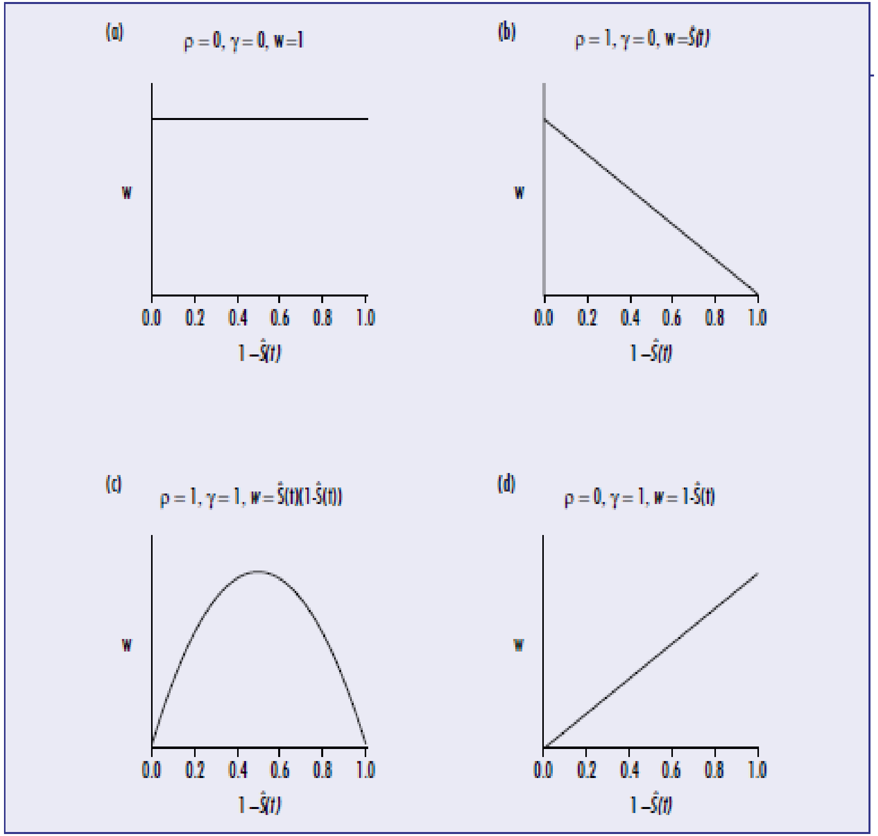
Test	Weight	Summary Statistics for treatment effect
log rank	1	Hazard ratio estimated by Cox regression with treatment as covariate
Gehan’s Wilcoxon	risk set	1. Average Hazard ratio with different choice of weights 2. Short term and long-term hazard ratios
generalized Wilcoxon	S_{KM}	
Tarone-Ware	Square root of risk set	
Fleming-Harrington	$S^{\rho}_{KM}(1- S_{KM})^{\Upsilon}$	
Supremum (Renyi)	supremum versions of the log-rank test	

Fleming and Harrington Weighted Log-rank (FH Log-rank)

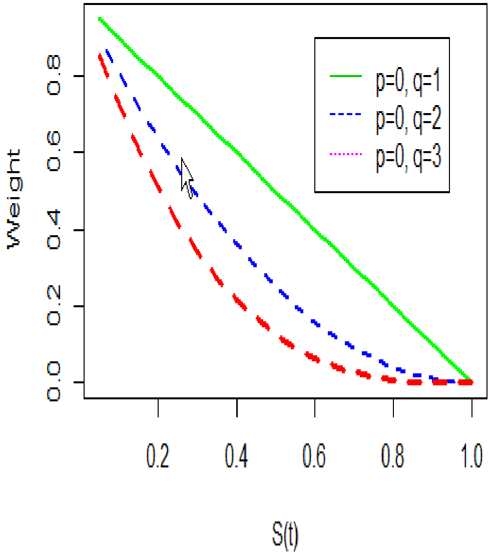
- Fleming and Harrington proposed the $G^{p,q}$ family based on the weight

$$W_n^{p,q}(s) = [\hat{S}_n(s)]^p[1 - \hat{S}_n(s)]^q$$

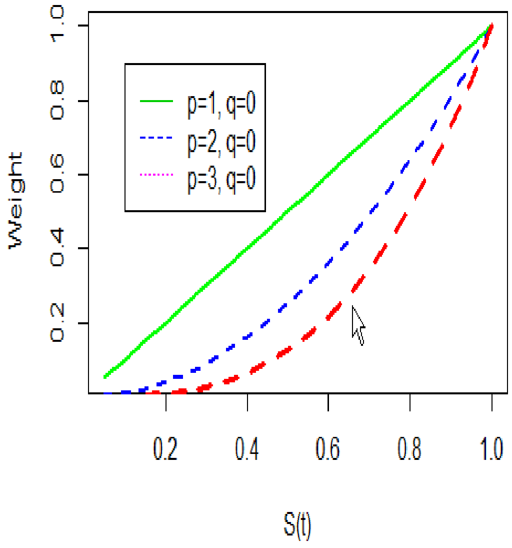
- Values of p and q implies
  - p> 0, q=0 : early difference
  - p=0, q>0 : late difference
  - p>0, q>0 : mid difference
  - P=0, q=0: log-rank test



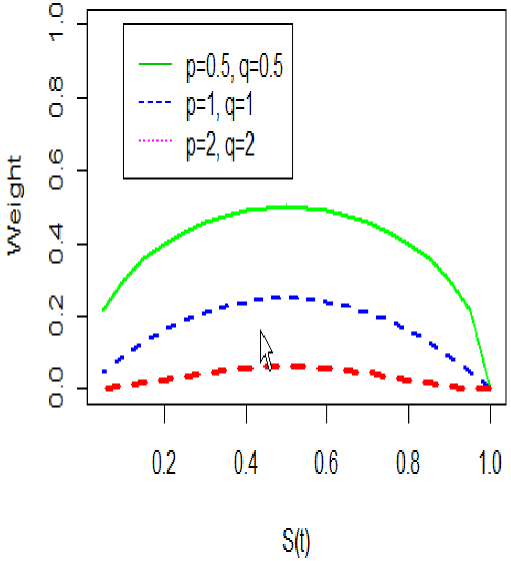
Combination FH's



Delayed Effect



Early Effect



Middle Effect

Example Revisited

Delay in effect: 2 months

- Re-analyzed the data using FH log-rank test with
- To cope with delayed effect p=0 and q= 1 were used (FH(0,1))
- Improved power with 263 events:

  - power **~ 74%**
- Average HR: weighted Cox regression
  - Estimate: **0.66 (0.46, 095)**

Delay in effect: 6 months

- Analysis using FH(0,1)
- The power loss is even severe
  - power: **~ 38%**
- Estimation of treatment also deteriorates
- The actual lag time has a significant impact on FH weighted lo rank test
- Can this be pre-specified for primary analysis?

Class of Alternatives: Can We Do Better with FH?

- The choice of a given weight depends heavily on the postulated alternative hypothesis
  - for example; lag time for treatment effect
  - Mis-specifying this alternative may imply a loss of power
- Combination test : a possible alternative
  - handles simultaneously a range of alternatives
  - provides robustness over the mis-specification
- Notable ones include
  - linear combinations of weighted statistics: Zucker (1992) and Lee (1996)
  - **maximum of several weighted logrank statistics**: Breslow et al. (1984), and Garès et al. (2015)

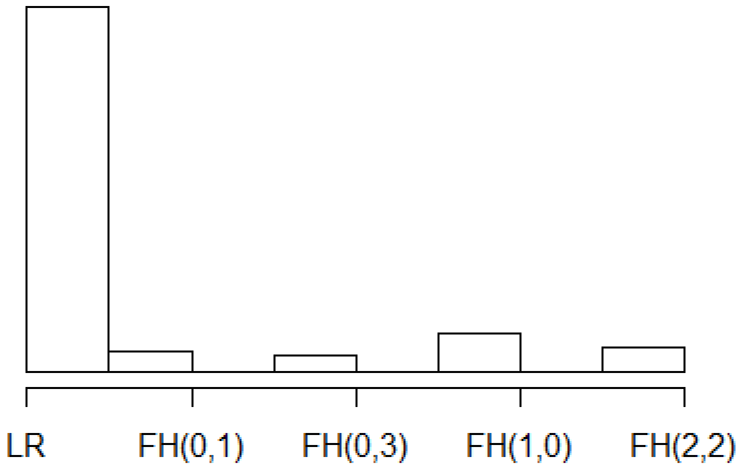
Combination of FH Statistics

- We have considered two combination possibilities
  - combination of $G^{0,0}$ and $G^{0,1}$: **Combo 1**
  - combination of $G^{0,0}$, $G^{0,1}$, $G^{0,3}$, $G^{3,0}$, $G^{2,2}$: **Combo 2**
- Max combination test: largest of the test statistics (smallest p-value)
- Requires multiplicity: Bonferroni-Holmes adjustment is used
  - maximal over adjusted p-values
- Estimation: weighted cox regression
  - weight (p,q) that provides maximal test statistics
- Provides great flexibility over different patterns

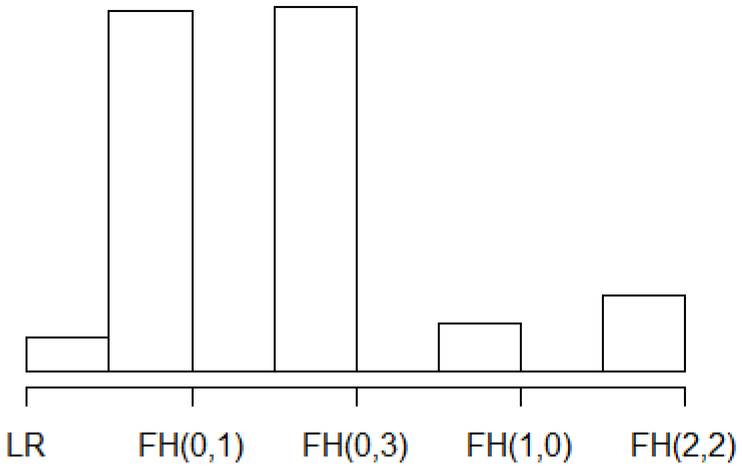
Simulation Study: Delayed Effect (N= 400, Event =320)

Control Median	True HR	Lag	LR	FH(0,1)	FH(1,0)	Combo1	Combo2
5	1	0	0.025	0.023	0.028	0.020	0.021
5	0.5	0	**0.94**	**0.92**	**0.90**	**0.89**	**0.89**
5	0.5	2	0.61	**0.90**	0.14	**0.89**	**0.89**
5	0.5	6	0.37	0.74	0.15	0.73	**0.89**
5, 7	1	0	0.025	0.022	0.026	0.021	0.019
5, 7	0.5	0	**0.93**	**0.90**	**0.90**	**0.89**	**0.89**
5, 7	0.5	2	0.57	**0.90**	0.14	**0.89**	**0.89**
5,7	0.5	6	0.34	0.74	0.15	0.73	**0.89**

Choice of Zmax Statistics



PH, HR = 0.5



lag= 6 month, HR = 0.5

Key Findings

- Further simulations are performed with
  - crossing hazard
  - diminishing treatment effect
- Results look promising
- Bias and coverage of average HR are also looked at
- A potential candidate for primary analysis if carefully applied

Treatment Effect Estimation: Role of Secondary Analysis

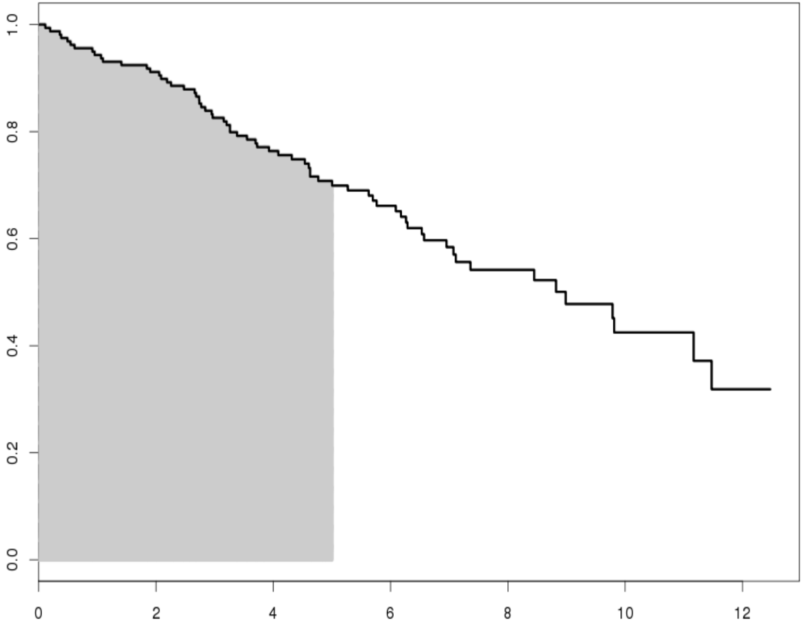
- While combination test and average ratio can construct primary measures of comparison, other supplemental measures are required if NPH is present
  - need sufficient information to understand the entire time profile
  - true HR is changing over time and therefore the interpretation of average HR is not intuitive
- Secondary or supplemental analysis plays an important role
  - requires clinically meaningful measures
- Work is still needed to understand how these information can be communicated in drug label
  - certain degree of pre-specification is required

Secondary Analysis: Need Careful Planning

- Graphical display
  - Kaplan-Meier plot
  - HR vs time plot
  - Area under the KM plot
- Milestone analysis
  - time points need to be pre-specified
  - very easy to understand by clinicians and patients
  - can be more powerful than average HR: need adequate number of events after separation
  - proper adjustment of variance is required

Restricted mean survival time (RMST)

- The shaded region (area under the survival curve) represents the RMST with a truncation time of 5 months
- The RMST calculated is 4.3
- The statistical interpretation would be “at 5 months, the mean survival of a patient is 4.3 months”
- The clinical interpretation would be “the life expectancy of the patient over the next 5 months is 4.3 months”



Other Analyses

- Piecewise hazard
  - powerful tool to understand the time profile of HR
- Other possible analysis includes
  - cox model with time dependent analysis
  - AFT
  - short term and long term HR’s
  - cure rate model if long term survivors are present
- Some examples in the next slides by digitalization of published KM curve

Example

Method	CM141	CM017
Cox regression	0.69 (0.52, 0.90)	0.66 (0.51, 0.87)
Cox with ln(t) as covariate	-0.33 (-0.52, -0.14)	-0.39 (-0.60, -0.18)
AFT model (Weibull)	0.39 (0.15, 0.64)	0.59 (0.31, 0.86)
Average HR	0.71 (0.54, 0.93)	**0.75** **(0.57, 1.00)**

RMST
τ= Max
 	
τ= 10mo
	
	1.27
	(1.07, 1.49)
	1.15
	(1.01,1.31)
		
		1.58
		(1.22, 2.04)
		1.33
		(1.11, 1.59)

YP model	L: 1.08 (0.25, 4.59) S: 0.39 (0.16, 0.91)	L: 1.02 (0.22, 4.78) S: 0.44 (0.18, 1.10)

Piecewise Analysis: CM 141

	Time Cut Point	Piecewise HR and 95% CI	P-value from Global Test
Equal Time	3m, 6m, 9m	• 1.06 (0.69, 1.62) • 0.44 (0.28, 0.70) • 0.61 (0.29, 1.29) • 0.55 (0.22, 1.37)	0.0040
Equal Percentile	1.875m, 3.375m, 5.6m	• 0.91 (0.52, 1.59) • 1.06 (0.60, 1.90) • 0.49 (0.29, 0.81) • 0.50 (0.29, 0.86)	0.0074
Eyeball check	3m, 6m, 9m	• 1.06 (0.69, 1.62) • 0.44 (0.28, 0.70) • 0.61 (0.29, 1.29) • 0.55 (0.22, 1.37)	0.0040

Piecewise Analysis: CM 17

	Time Cut Point	Piecewise HR and 95% CI	P-value from Global Test
CoxPH	n/a	• 0.66 (0.51, 0.87)	0.0025
Equal Time	3m, 6m, 9m	• 0.87 (0.62, 1.22) • 0.44 (0.25, 0.76) • 0.31 (0.15, 0.66) • 1.59 (0.35, 7.19)	0.0008
Equal Percentile	1.5m, 2.15m, 4.8m	• 0.82 (0.49, 1.39) • 1.18 (0.70, 1.99) • 0.38 (0.21, 0.67) • 0.51 (0.30, 0.86	0.0009
Eyeball check	3m, 6m, 9m	• 0.87 (0.62, 1.22) • 0.44 (0.25, 0.76) • 0.31 (0.15, 0.66) • 1.59 (0.35, 7.19)	0.0008

Design Aspects

- The initial estimate can be based on
  - log-rank test with average HR
  - piecewise exponential approximation
- Planning of follow up time is key
- Interim analysis
  - Early futility can be problematic
- Evaluate power and sample size considerations across different scenarios
- Make recommendations that may be acceptable
  - Noting implied limitations if design assumptions are in the wrong scenario

Cross Pharma NPH Working Group

- Address the issue of NPH for design, analysis and interpretation
  - **Across Oncology/ImmunoOncology:** NPH has been seen in various settings including Oncology trials and more recently the IO trials
  - **Focus on Phase III trials**: the team will focus on Phase III/regulatory trial setting
  - **Impact on Payer Analyses:** not in the initial scope; to follow later
  - **Audience:** initial audience- statisticians, clinicians, regulatory

Group Structure

- **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), Julie Cong (B&I), Satrajit Roychoudhury (Novartis), Tianle Hu (Lilly)**
    - **Endpoint Workstream: Jane Qian (Abbvie), Dominik Heinzmann (Roche)**
    - **Simulation Team: Julie Cong (B&I)**

  – **Organizations represented in the Working Group**

    - **AZ, BMS, Merck, B&I, Novartis, Lilly, Abbvie, Roche, Bayer, Janssen, Takeda, Amgen, Pfizer, GSK, Celgene, and FDA**

Ongoing Activities

- – **Initial kick off meeting:** October 2016
- – **Work stream formed, leaders in place and fully kicked off**: January 2017
- – **Face to Face midpoint meeting**: ASCO 2017
- – **Goals:** Conference February 2018 and Guidance/White Papers Spring 2018

Concluding Remarks

- A careful consideration of NPH is required for in design and analysis phase
- Combination of FH weighted log rank test provides an robust alternative
  - handles different NPH pattern simultaneously
- Average can be used for primary analysis, but supplemental analyses are required
  - Need clear understanding of clinical benefit by non-statisticians
- A cross phama company collaboration group is working to understand the practical issues and provide solutions
  - planning joint workshop with FDA early next year

Acknowledgement

- Keaven Anderson
- Tianle Hu
- Renee Iacona
- Tai-Tsang Chen
- Valuable inputs from other members of the working group

Thank You