# Statistical challenges in immunotherapy trials

*European Organization for Research and Treatment of Cancer's*

## Perspectives

Michal Kicinski, PhD

Nathan Touati, MS

EORTC

*The future of cancer therapy*

# Cancer Immunotherapy

Major focus area for cancer treatment

- Growing excitement about these agents the past few years

***Immunotherapies*** *targeting the immune system*

*vs*

***Chemotherapy*** *and* ***targeted therapies*** *targeting directly the tumor*

- The immune and anti-tumor response to immunotherapies is <u>dynamic</u>

Innovative mechanism of action poses challenges for the classical methodology for trial design and analysis

- Challenges are both **<u>clinical</u>** and **<u>statistical</u>**

# New challenges – Response assessment (**iRECIST**)

Response assessment by RECIST criteria is globally accepted

- Issue with immunotherapy: other response patterns have been observed, leading to PD diagnosis and therefore discontinuation of experimental treatment

*Alternatives*:


2009: immune-related response criteria (irRC) based on WHO criteria


2017: Newly proposed consensus-based guidelines: **iRECIST**

*Source: Seymour 2017, "iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics", Lancet Oncol*

# New challenges – Response assessment (**iRECIST**)

Recommendations:

- **Phase 3 trials**: incorporate both RECIST 1.1 and iRECIST
  - **But** RECIST 1.1 should continue to be used to define the primary efficacy outcomes (*when PFS, disease progression, BOR…*)

- **Early-phase trials**: can consider using iRECIST as primary criteria

Next step:

- Validation of the efficacy of iRECIST with the creation and analysis of a warehouse of data from immunotherapeutic trials

# New challenges – Safety & Endpoint

## Caution: Safety with immunotherapies

Careful **monitoring of immune-related adverse events (irAEs)** is required during both trial and long-term follow-up

- majority of immune-mediated reactions occur during the initial stages of the treatment

## What is the most appropriate endpoint?

**Overall survival** (OS) remains the gold standard

- *Approval of ipilimumab, nivolumab and pembrolizumab in advanced lung cancer and melanoma was based on OS*

But OS comparisons can be confounded by:

- crossover within a trial
- subsequent treatments
- competing non-cancer related events

*Alternative*: immune-related PFS (PFS by irRC) exist but is not yet commonly used

# Main statistical issue: Non-PH treatment effects

General assumption in trials: Proportional-Hazards treatment effect

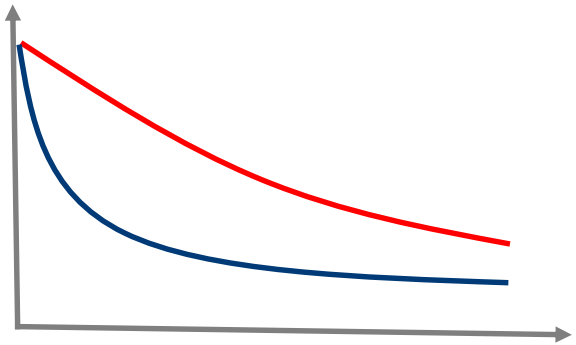
- Sample size based on log-rank test
- HR estimate based on Cox PH model

In immunotherapy, treatment effect may depends on time

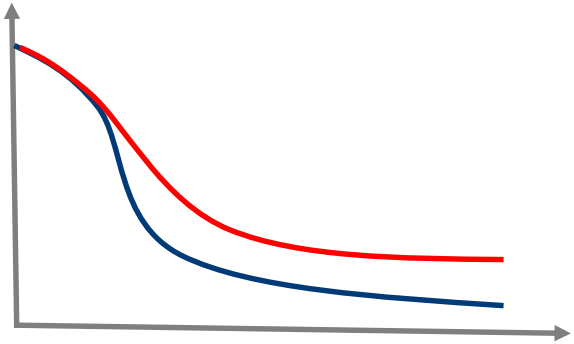
- Caused by late immune response, short period of treatment administration…

Various patterns of non-PH effects:

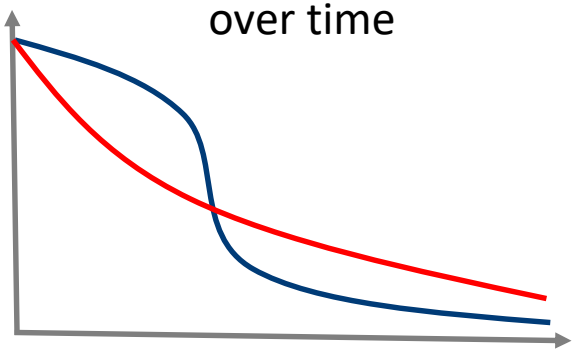
Early treatment effect



Delayed treatment effect

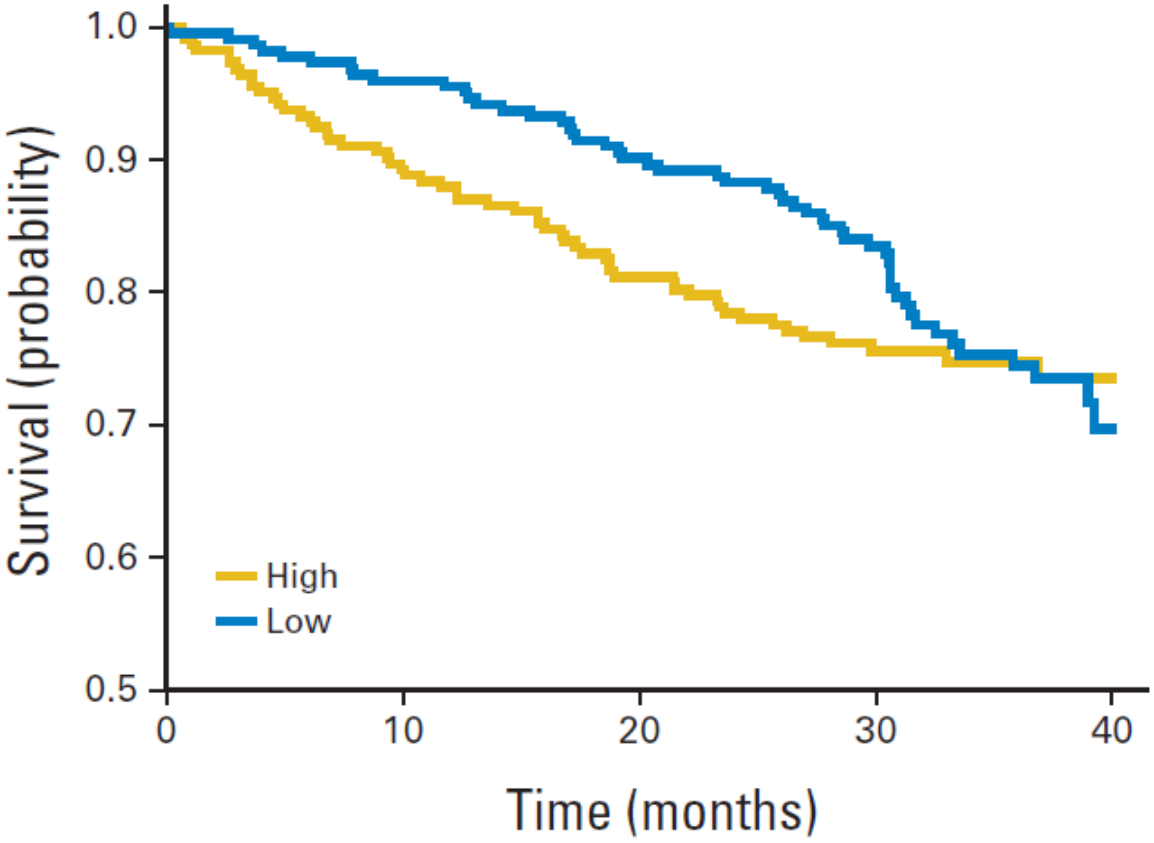


Reverse treatment effect over time



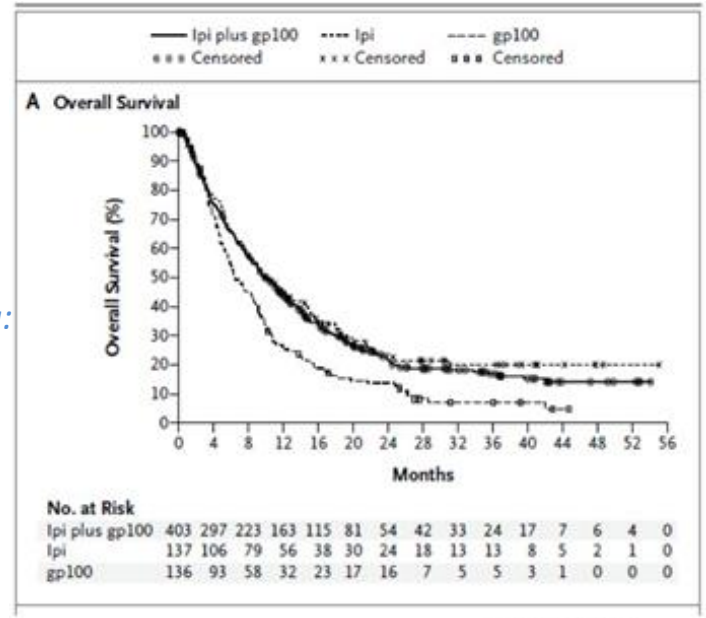
# Examples of non-PH patterns (1) – **Early effect**

*Eastern Cooperative Oncology Group E4A03 study*



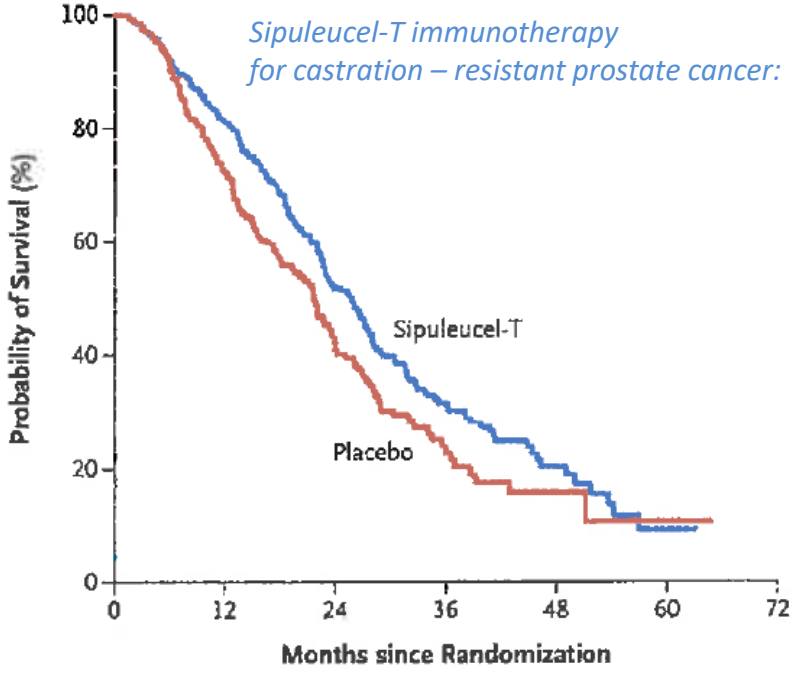
# Examples of non-PH patterns (2) – **Delayed effect**

*Ipilimumab in Patients with Metastatic Melanoma:*

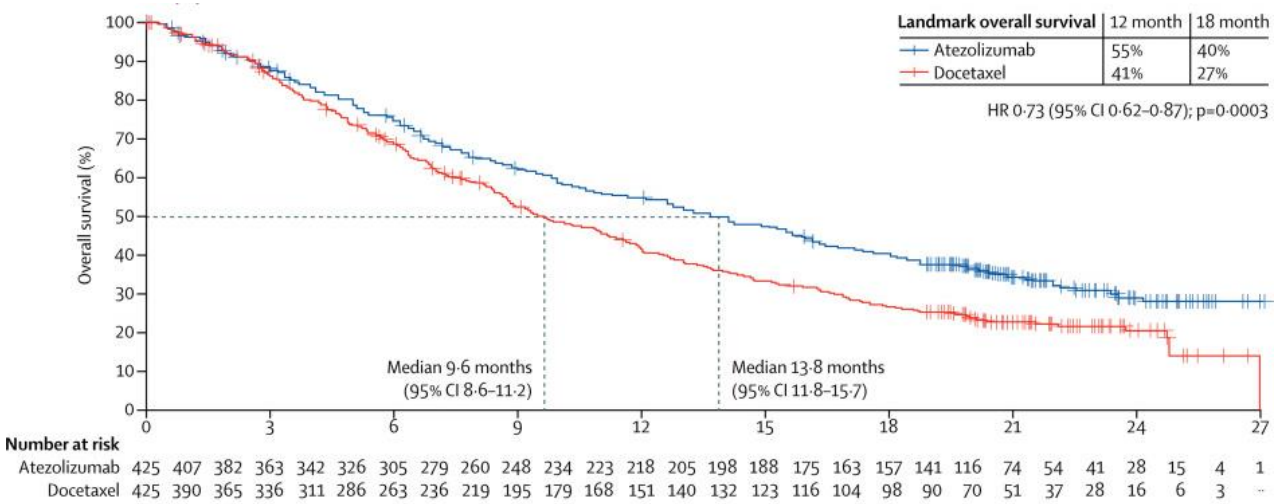



Hodi et al. NEJM 2010,363:8

*Sipuleucel-T immunotherapy for castration – resistant prostate cancer:*



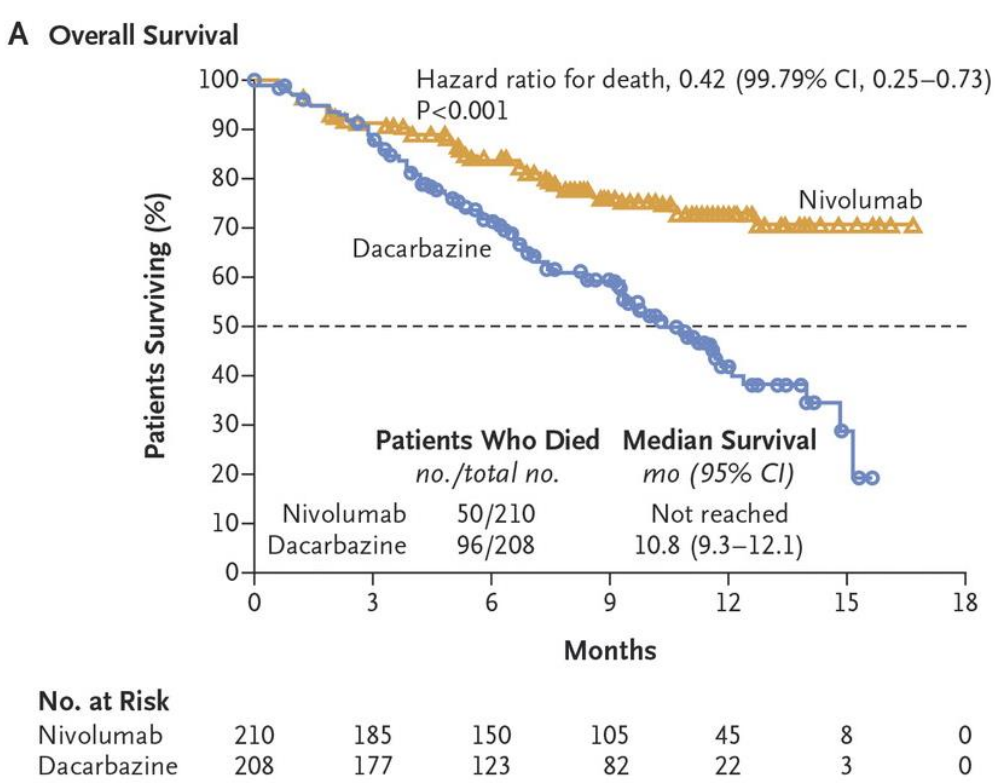

*Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre*

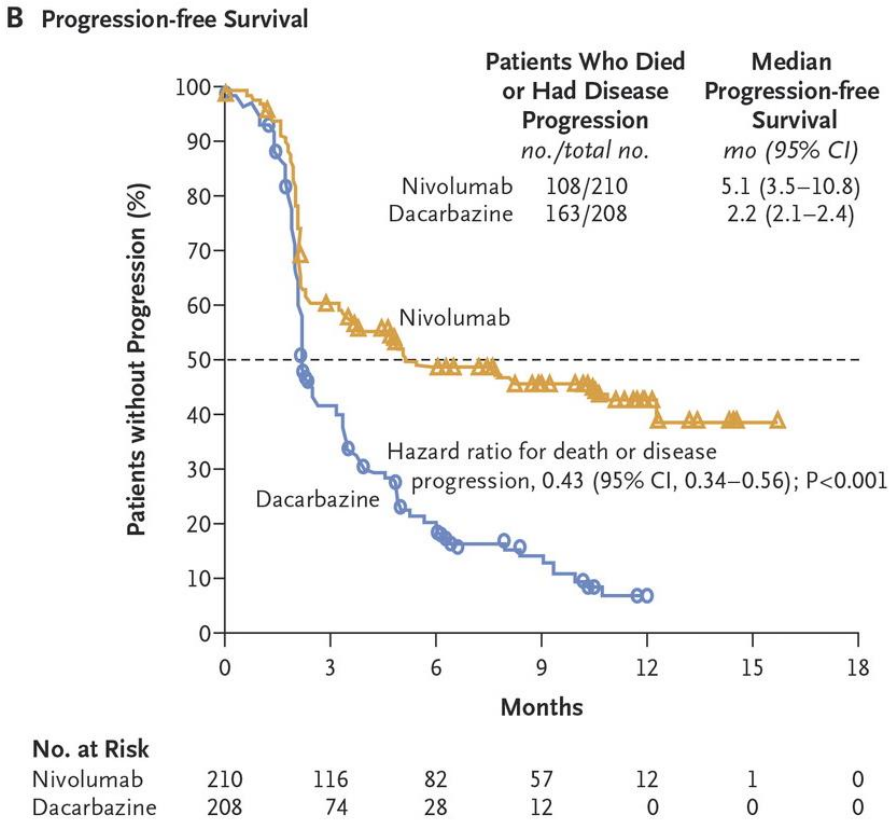



EORTC

# Examples of non-PH patterns (2) – **Delayed effect**

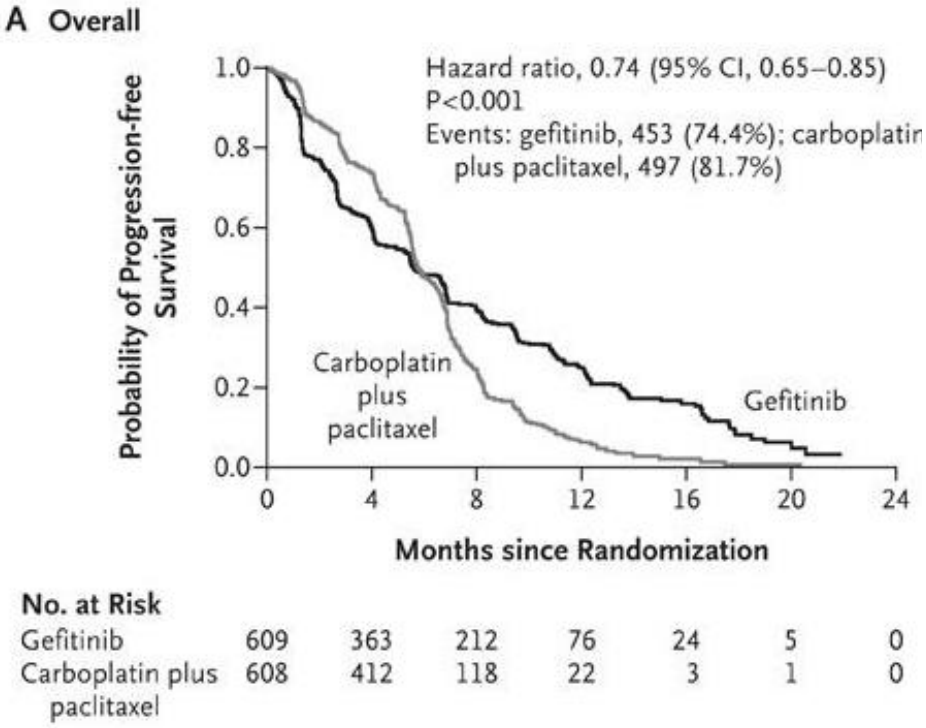
*Nivolumab versus Dacarbazine in Previously Untreated Melanoma without BRAF Mutation*



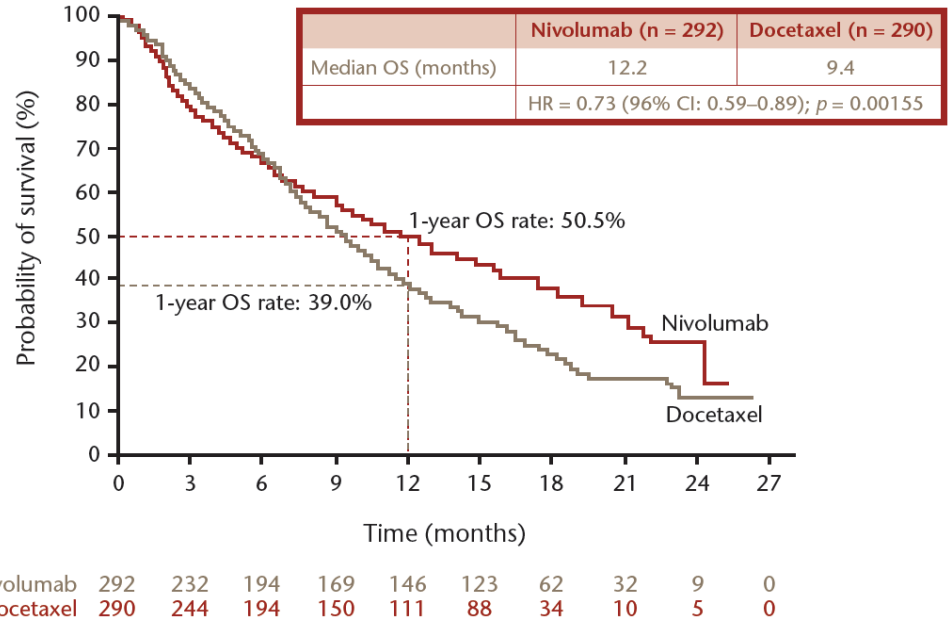



# Examples of non-PH patterns(3) – **Reverse effect over time**

*Gefitinib or Carboplatin-Paclitaxel in Pulmonary Adenocarcinoma*




*Checkmate 057 Nivolumab vs Docetaxel in NonSquamous Non-Small-Cell Lung Cancer*



# Log-rank test

- The **most powerful non-parametric test** to compare survival functions **under PH**
- Equivalent to the score test for HR from the Cox model: **test-estimation coherency**

**Standard practice**

# If non-PH

- Log-rank test may not be the most powerful non-parametric test
- The interpretation of the corresponding treatment effect (HR) complicated
  - Is not a simple average of the hazard ratios over time
  - HR depends on the censoring distribution, which is study-specific

**Standard practice: always optimal?**

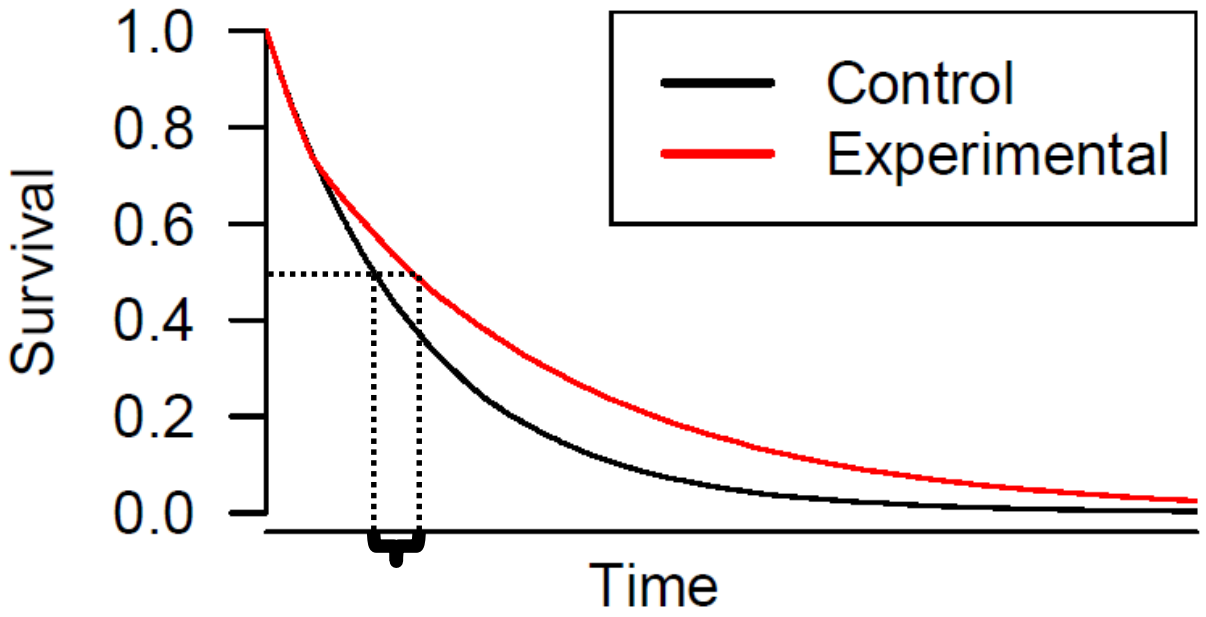
# Wish list

- Proper control of type I error probability
- Substantial power gain compared to the standard practice when hazards non-proportional
- Limited power loss when hazards proportional
- Test-estimation coherency
- Simple and meaningful interpretation
- Possibility to condition on stratification factors

# What else if not HR

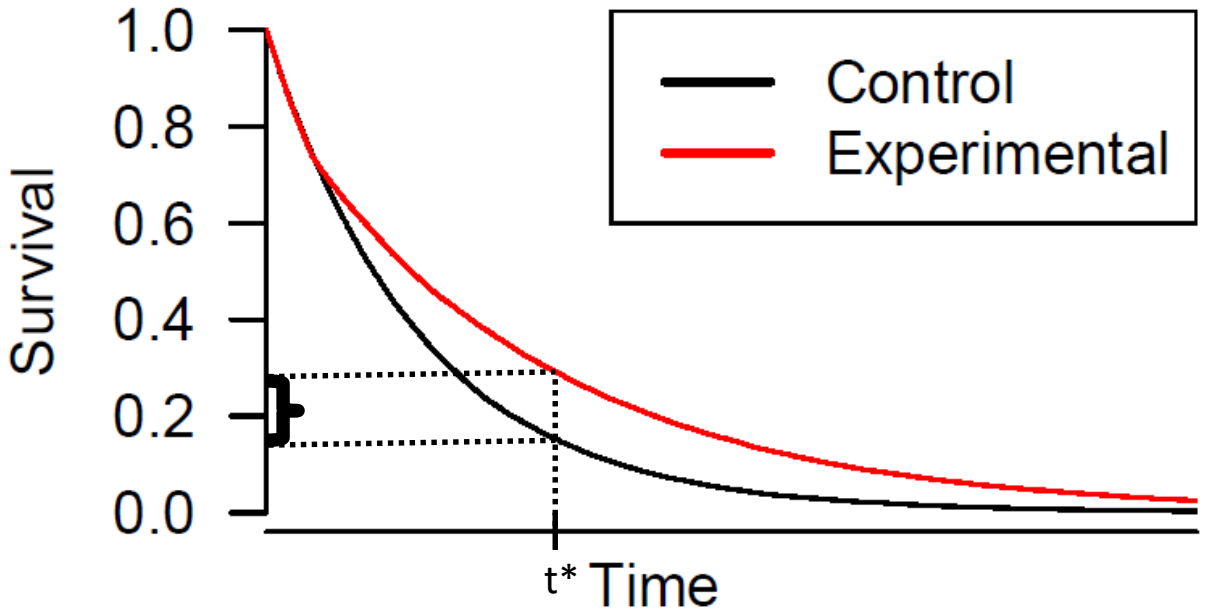
- Median survival time (or other percentile)
- Survival time at certain time point
- Restricted mean survival time

# Median survival time




- Less technical interpretation than HR
- Inference based on the difference
- Ignores what happens after the median has been reached (efficiency loss)
- For PFS, depends on the time timing of the scans
- Why this percentile?

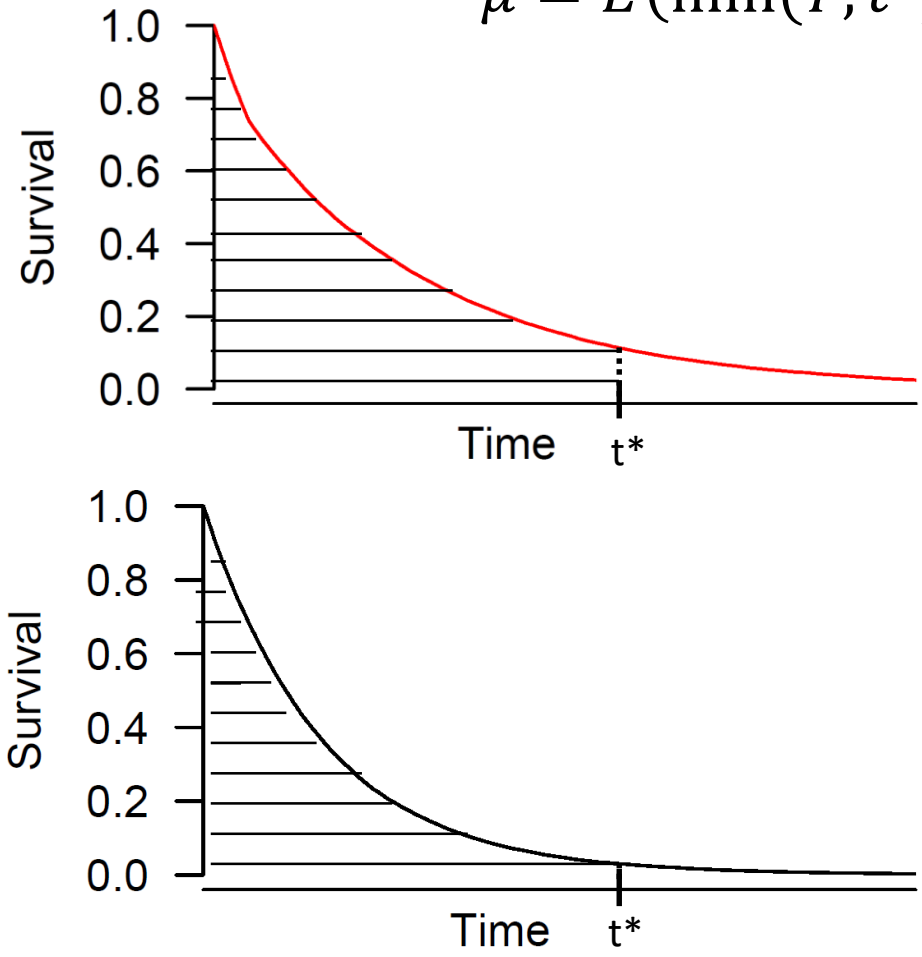
# Survival time at certain time point




- A comparison based on the KM estimates
- Appealing interpretation
- May suffer from lack of efficiency
  - Ignores what happens after the chosen time point has been reached
  - Not all patients are recruited at the same time so some events are excluded
- The choice of the time point subjective

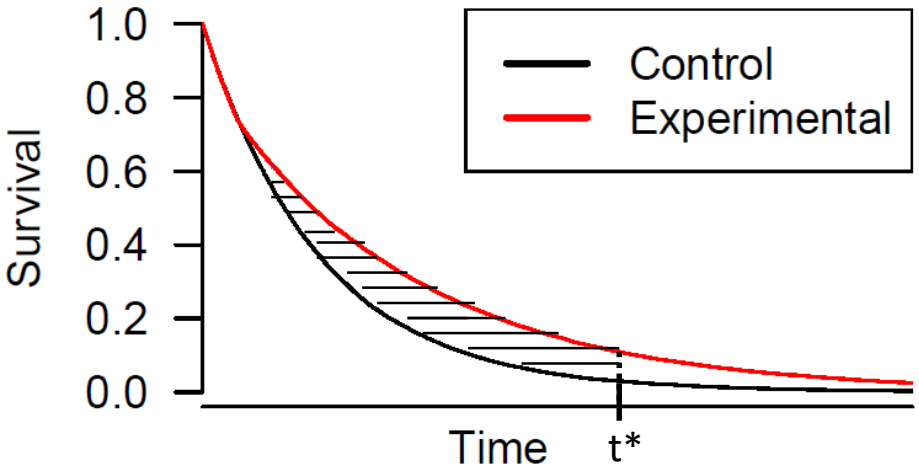
# Restricted mean survival time

$$\mu = E(\min(T, t^*)) = \int_0^{t^*} S(t)dt$$




- Several methods of estimation exist
- Better options than direct integration using KM estimates exist, especially for small sample sizes

# Restricted mean survival time




**Simple and meaningful interpretation irrespective of the effect pattern**

- Comparison usually based on the difference (proportion also possible)

$$\hat{\Delta} = \widehat{\mu_1} - \widehat{\mu_2}$$

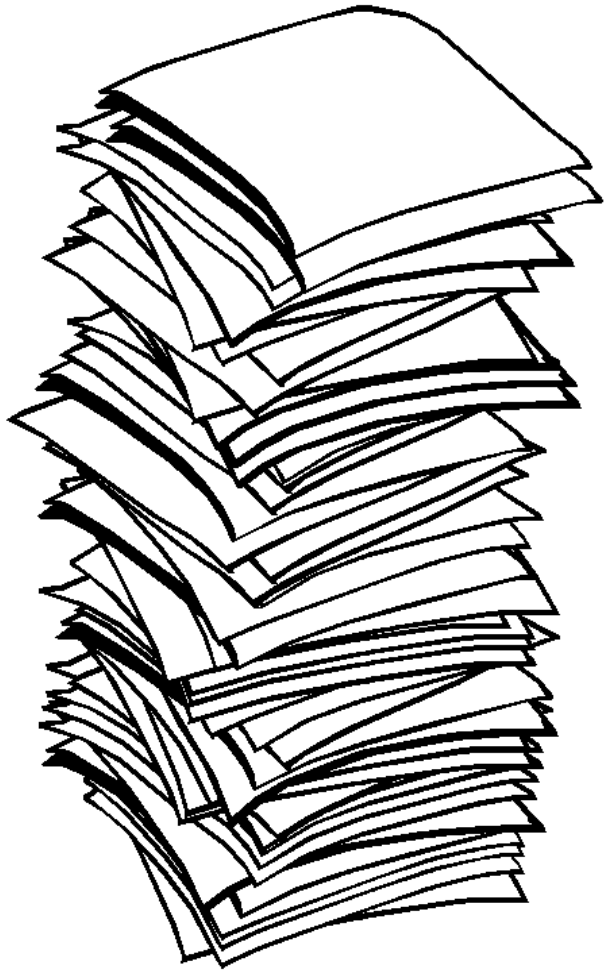
- Statistical inference using:

$$\frac{\hat{\Delta}}{\sqrt{var(\hat{\Delta})}} \rightarrow N(0,1), \text{ with } var(\widehat{\mu_j}) \text{ estimated using the delta method}$$

- Incorporation of covariates possible
- For small sample sizes, a permutation test should be used

# Other alternatives to the log-rank test

- Other tests based on event rates
- Kaplan-Meier test statistics
- Adaptive tests based on restricted mean survival time
- Combinations of two approaches



# Other tests based on event rates

- Classical weighted log-rank tests (for two groups)

$$Z = \sum w_{t_i}(o_{t_i}^1 - e_{t_i}^1)$$

$$\frac{Z^2}{var(Z)} \sim \chi^2(1)$$

- e.g., Fleming-Harrington class of weight functions:

$$w_{t_i} = [\hat{S}(t-)]^q[1 - \hat{S}(t-)]^\gamma, q \geq 0 \text{ and } \gamma \geq 0$$

  - $q = 0$ and $\gamma = 1$: a popular test emphasizing late differences
  - $q = \gamma = 0$ gives the log-rank test

- Assign $w_1$ to early event times and $w_2$ to late event times (Xu et al. Stat Med 2016)
- Adaptively weighted log-rank test (Yang and Prentice Biometrics 2010)

# Kaplan-Meier test statistics

- Test statistics based on a weighted average of the difference between the KM survival estimates at different time points
- Higher weight given to time points with bigger differences and larger number of patients at risk
- Several approaches to estimate the weights have been proposed (Shen and Cai Biometrics 2001; Uno et al. Stat Med 2015)

# Combinations of two approaches

- Two log-rank tests (Sit et al. Stat Med 2016):
  - Non-inferiority test for the whole study period:

$$H_0^{(1)}: {}^{\lambda_1}/_{\lambda_2} \geq a_1 (> 1) \text{ for } t \geq 0$$

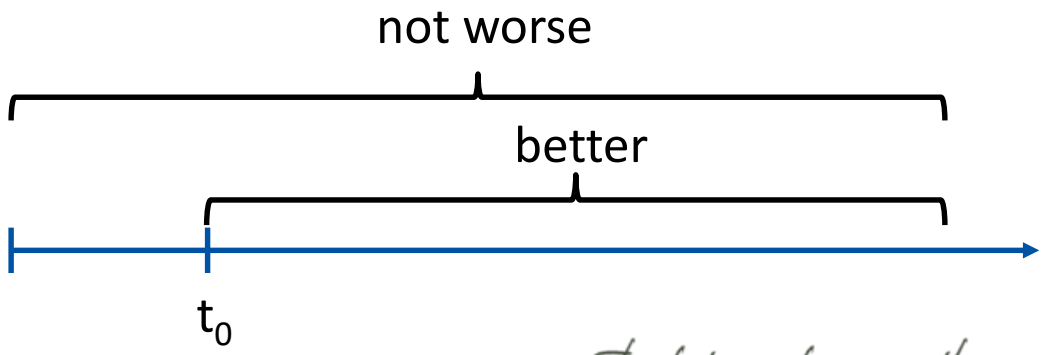
  - Superiority test for the period from time $t_0$:

$$H_0^{(2)}: {}^{\lambda_1}/_{\lambda_2} \geq a_2 (\leq 1) \text{ for } t \geq t_0$$

- Both null hypothesis need to be rejected to claim superiority

$$H_0: H_0^{(1)} \cup H_0^{(2)}$$

$$H_1: H_1^{(1)} \cap H_1^{(2)}$$



# Combinations of two approaches

- A two-stage procedure of Qiu and Sheng (Statist Soc B 2008)
  - Stage I: log-rank test
  - Stage II (only in case stage I does not reject $H_0$): a test to distinguish cases when the hazard rates are identical and cross each other
  - Test in stage II independent of the log-rank test, so no correction for multiplicity needed

# Combinations of two approaches

- Augmented log-rank test (Royston and Parmar BMC Med Res Meth 2016)
  - Calculate log-rank test p-value $p_{L\text{-}R}$
  - Calculate the p-value of the permutation test for RMST $p_{RMST}$
  - Take the minimum $p_{min}=\min(p_{L\text{-}R}, p_{RMST})$
  - Compare $p_{min}$ to the empirical distribution of $P_{min}$ under $H_0$

# Properties of methods to test survival differences

- Many methods (e.g., Yang and Prentice Biometrics 2010; Uno et al. Stat Med 2015) characterized by:
  - Substantial or even impressive power gain when the treatment effective and hazards non-proportional
  - Moderate or small power loss under PH
  - Correct type I error rate for a scenario of equal survival distributions, i.e., $S_1(t) = S_2(t)$ for all $t$

# $H_0$ matters

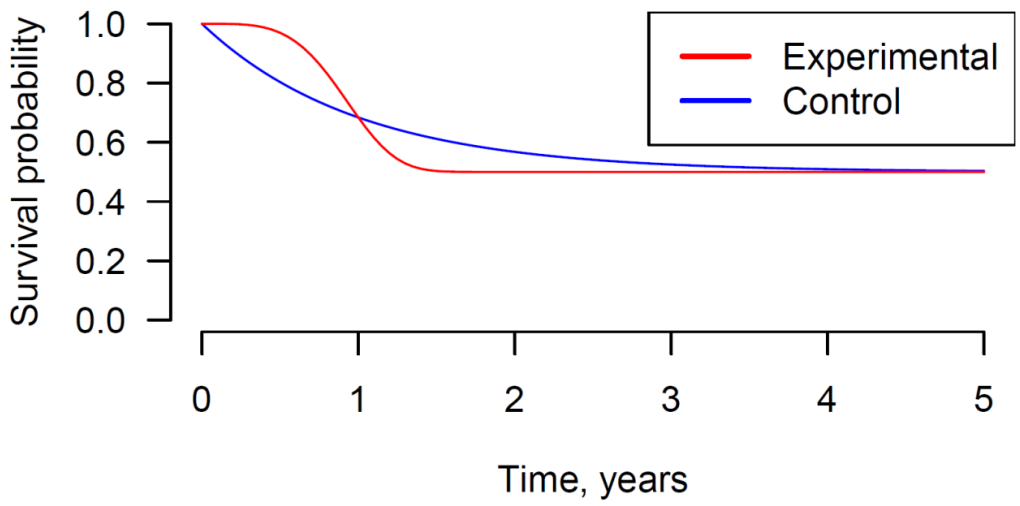
- Many proposed approaches (e.g., classical weighted log rank tests; Xu et al. Stat Med 2016; Yang and Prentice Biometrics 2010; Shen and Cai Biometrics 2001; Uno et al. Stat Med 2015; Qiu and Sheng Statist Soc B 2008; …) test the following hypothesis:

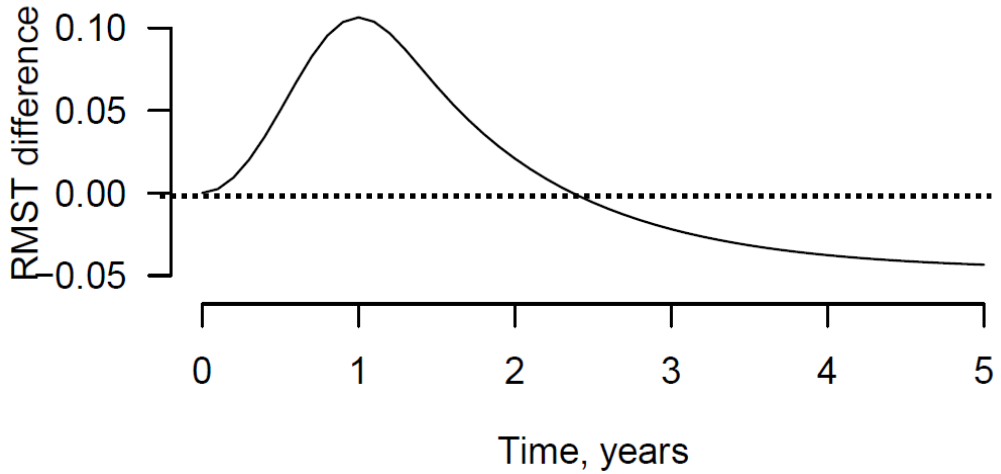
$$H_0: S_1(t) = S_2(t) \text{ for } \textbf{all } t$$

$$H_1: S_1(t) \neq S_2(t) \text{ for } \textbf{some } t$$

**TRUE $H_1$ DOES NOT IMPLY BETTER SURVIVAL IN ONE ARM!**

# Illustrative example

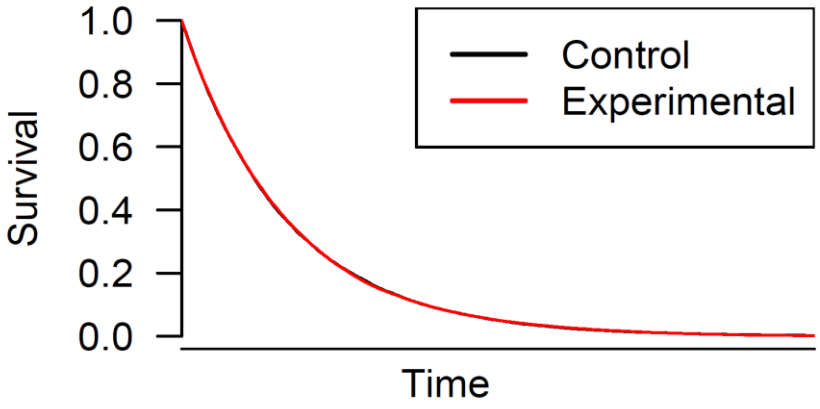


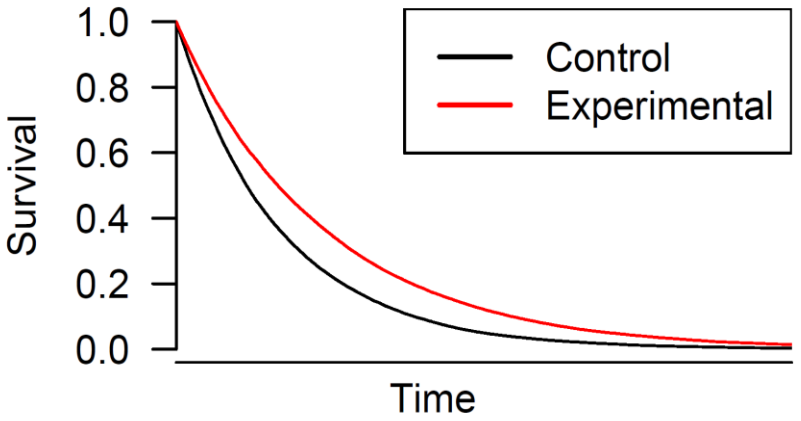

- Long-term survival probability of 0.5 in both arms
- Hazard first lower than higher in the experimental arm
- Mean survival time longer in the control arm
- Survival in the experimental arm better than in the control arm for some t
- The hazard function in the experimental arm lower than in the control arm for some t

# Simulation study

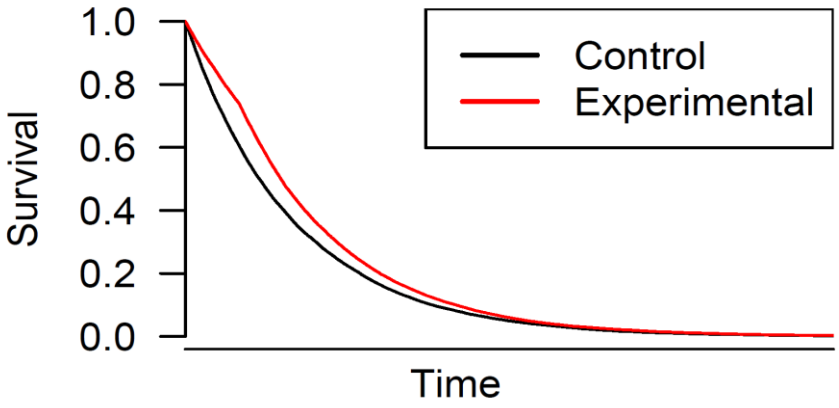
1) No effect



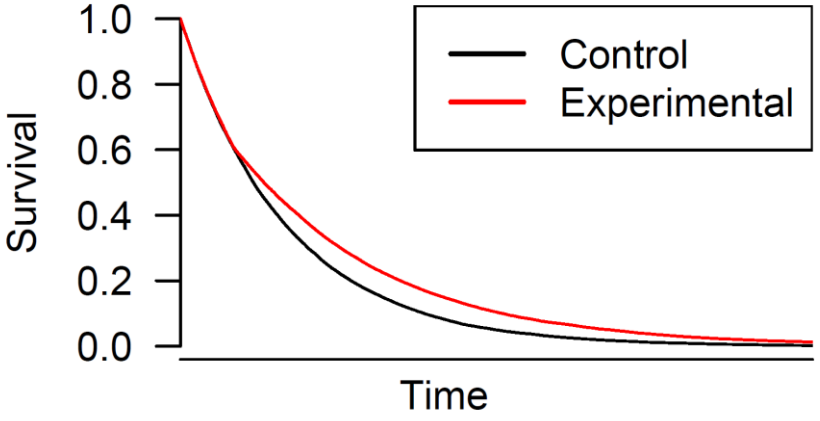

2) PH (HR=0.7)




3) Early benefit (HR=0.6 the first 6 months)

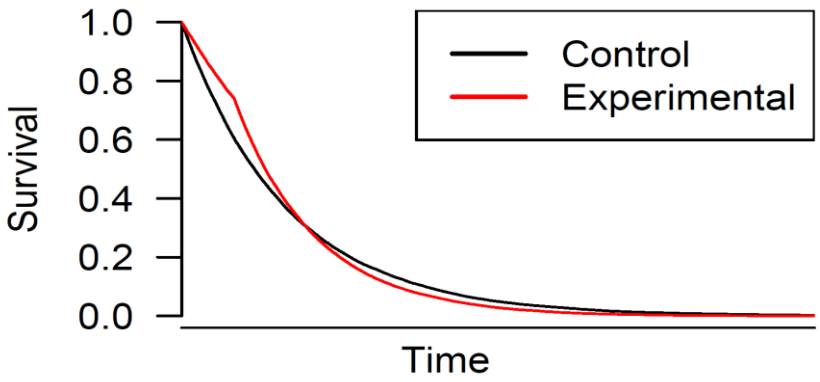



4) Late benefit (HR=0.7 from month 7)

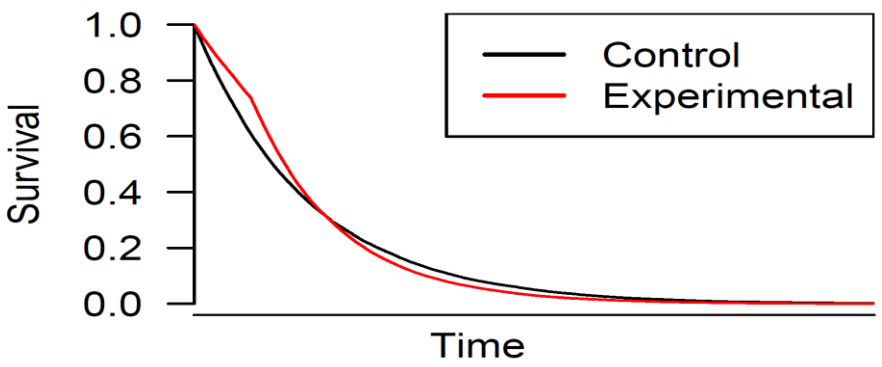


# Simulation study

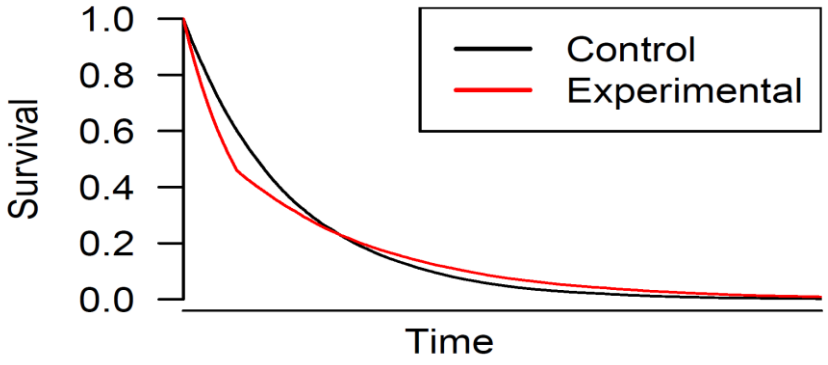
5) Early benefit, MSTD=0



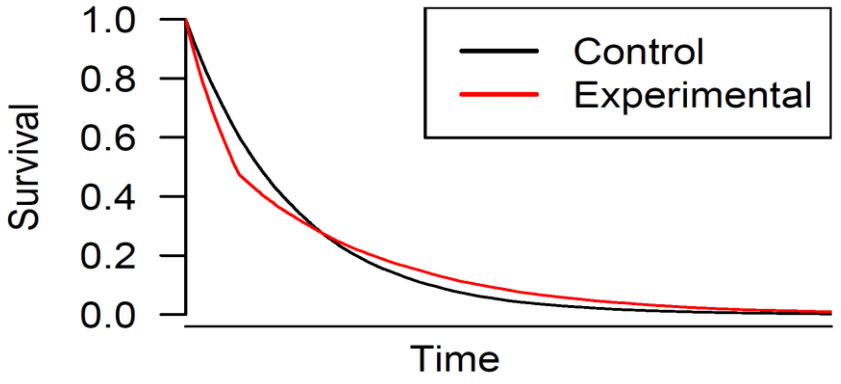

6) Early benefit, LRTS=0




7) Late benefit, MSTD=0




8) Late benefit, LRTS=0




MSTD: mean survival time difference LRTS: true likelihood ratio test statistic

# Simulation results

**Probability of rejecting $H_0$ for different tests.**

| Effect | Log-rank | RMSTD | Adaptive log-rank (Yang and Prentice 2010) |
|---|---|---|---|
| No effect | 0.05 | 0.04 | 0.06 |
| PH | 0.79 | 0.76 | 0.78 |
| Early benefit | 0.25 | 0.26 | 0.32 |
| Late benefit | 0.36 | 0.33 | 0.42 |
| Early benefit, MSTD=0 | 0.05 | 0.04 | 0.16 |
| Early benefit, LRTS=0 | 0.04 | 0.05 | 0.16 |
| Late benefit, MSTD=0 | 0.06 | 0.05 | 0.24 |
| Late benefit, LRTS=0 | 0.05 | 0.05 | 0.21 |

PH: proportional hazards; MSTD: mean survival time difference;
LRTS: true likelihood ratio test statistic

No censoring, no prognostic factors, a simple randomization, n=247 (corresponding to a power of 0.8 to detect HR=0.7 by the log-rank test), B=2000

# Simulation results

- The adaptively weighted log-rank test frequently rejects $H_0$ when there is no overall survival benefit in terms of the mean survival time or the event rate averaged over the whole follow-up
- A significant power gain related to the use of a test based on RMSTD compared to the log-rank test when the hazards are non-proportional not evident at all for the considered scenario

# Conclusions – testing procedure choice (1)

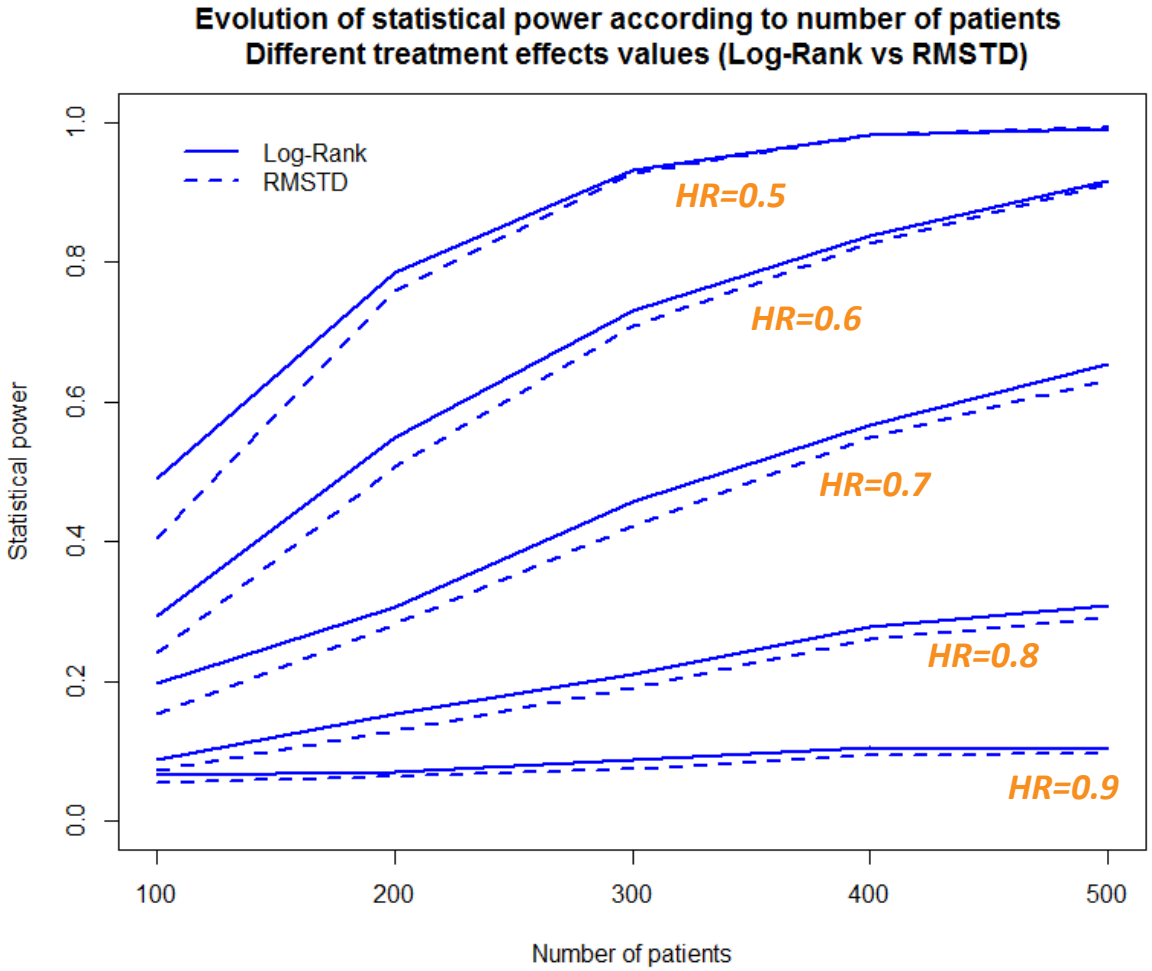
- When non-PH
  - The interpretation of the HR complicated
  - Log-rank test may not be the most powerful test
- The impressive power gain of some approaches comes with a cost of testing a wrong $H_0$ (be careful!)
- For methods based on the correct $H_0$, power gain (if any!) compared to the log-rank test under non-PH may be small

# Conclusions – testing procedure choice (2)

- Restricted mean survival time difference is an attractive alternative to HR when hazards are likely non-proportional
  - Proper control of type I error probability
  - Test-estimation coherency
  - Simple and meaningful interpretation
  - Possibility to condition on stratification factors
  - **Some** power gain for **some** scenarios

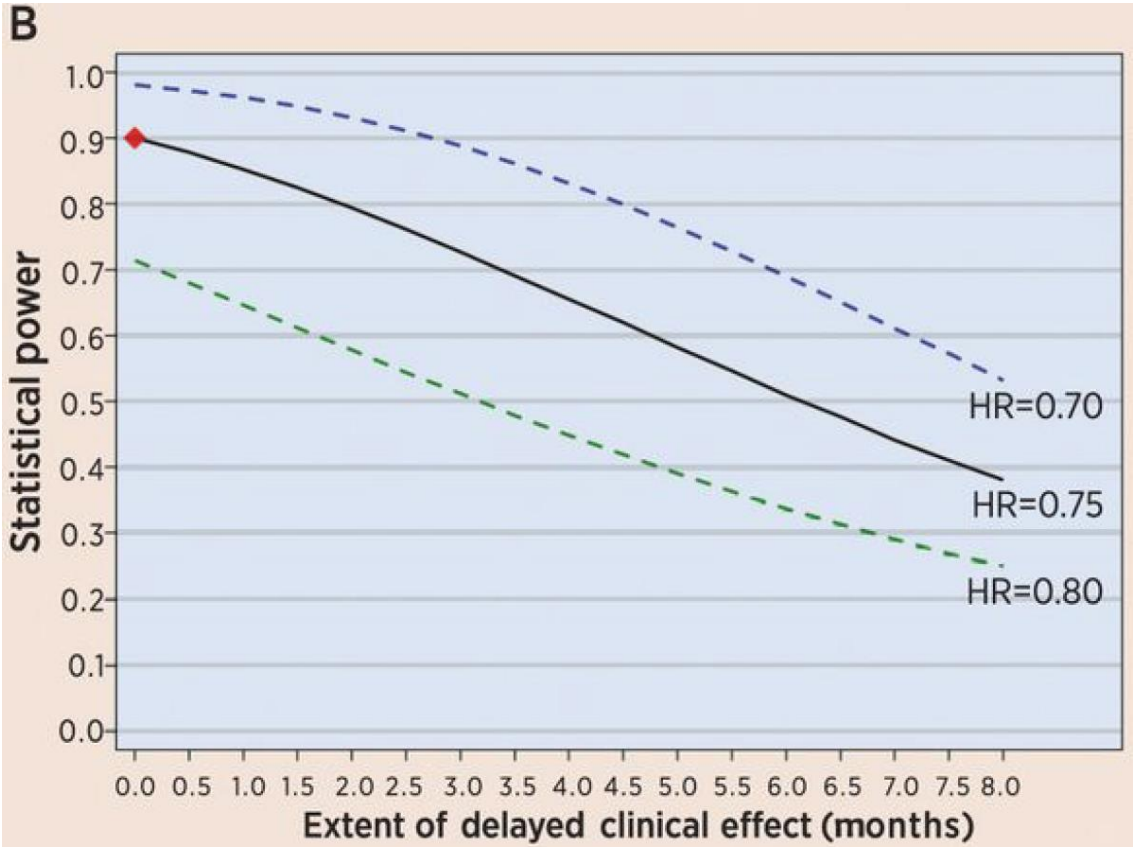
# Consequences of non-PH effects on study design

*Nb Sim= **2000**,*
*no censoring*
*Delay timing = **6 months***




- Statistical power is dependent of magnitude of treatment effect
- Small differences between Log-Rand and RMSTD-based test tend to shrink with higher number of patients

# Consequences of non-PH effects on study design




- Statistical power is dependent of both **timing of delayed separation** and **magnitude of treatment effect**

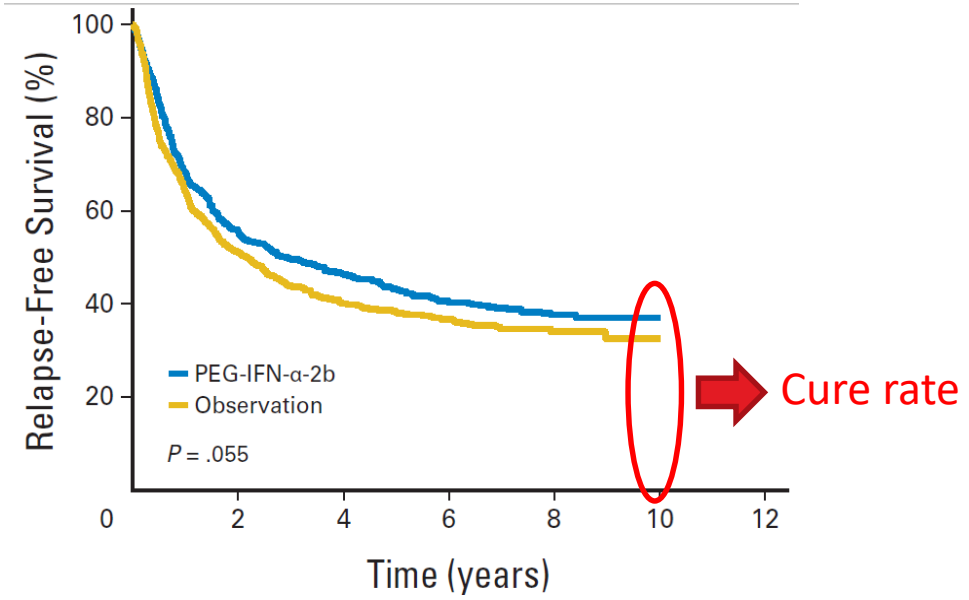
*Sources: Mick 2015, "Statistical Challenges in the Design of Late-Stage Cancer Immunotherapy Studies", Cancer Immunology Research*

# Trial issues – Long term survival

Studies are usually designed on exponential distribution assumption
→ survival curves will drop down to zero survival probability

*New setting with immunotherapies*: a subset of patients are <u>expected to be cured</u>

*Example: Pegylated Interferon alfa-2b (Sylatron):*
*Relapse-Free Survival – Adjuvant Melanoma*

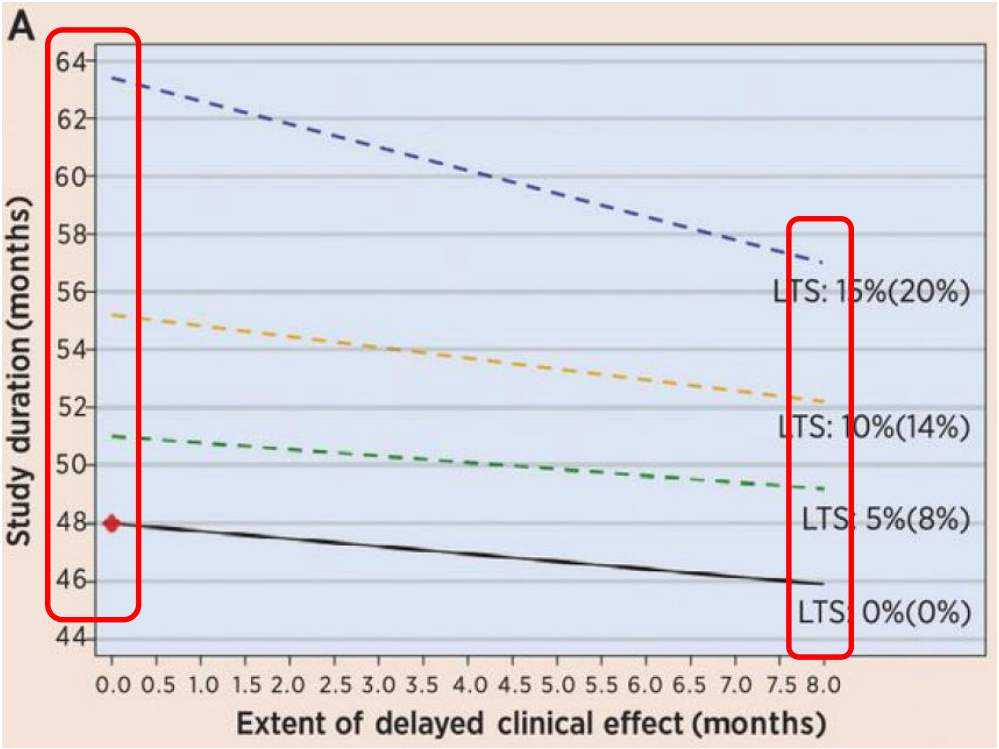



- Introduction of the **<u>Cure rate</u>**: % of long-term survival patients among all patients

*<u>Sources</u>: Chen 2013, "Statistical issues and challenges in immuno-oncology", Journal for ImmunoTherapy of Cancer*
*Mick 2015, "Statistical Challenges in the Design of Late-Stage Cancer Immunotherapy Studies", Cancer Immunology Research*

# Consequences of long-term survival

*LTS: Long-term Survival Control (Treatment), HR=0.75 after separation, 512 events*




***PH**: prolongation from 3 to 16 months for cure rate from 5 to 15%*

***Non-PH**: from 3 to 11 months for cure rate from 5% to 15%*

***Observation***: higher cure rate results in a longer time to reach the pre-specified number of events

The presence of long-term survival would lead to a **prolongation of trial duration**

In reality, cure may occur only in the treatment arm → *over-powered study*

*Sources: Mick 2015, "Statistical Challenges in the Design of Late-Stage Cancer Immunotherapy Studies", Cancer Immunology Research*

# Interim analysis – Impact of non-PH/LT data

*Accrual rate=20 patients per months; Interim analysis at the information fraction of 50%; design to detect HR=0.75, 2-sided type I error; delayed treatment effect = 3 months; Cure rate = 10% (control) vs 18% (treatment); O'Brien-Flemings boundaries*

**Interim stopping probability with long-term survival *(10% vs 18%)* and delayed clinical effect *(3 months)***

| | Standard PH model | LT survival | Delay | LT survival & Delay |
|---|---|---|---|---|
| Interim sample size | 520 | 540 | 480 | 500 |
| Number of events | 256 | 256 | 256 | 256 |
| Stopping probability (**superiority**) | 0.25 | 0.25 | 0.06 | 0.06 |
| Stopping probability (**futility**) | 0.01 | 0.01 | 0.08 | 0.08 |

vs vs vs

Long-term survival: **no apparent impact**

Delayed treatment effect: **high impact**

↗ False Negative Rate *(futility)*

↘ True Positive Rate *(superiority)*

*Sources: Chen 2013, "Statistical issues and challenges in immuno-oncology", Journal for ImmunoTherapy of Cancer*

# Interim analysis - Warning

Interim analyses should be implemented with **caution** in immuno-oncology trials

- Waste of resources and/or false conclusions
- Envisage the optimal time-point accounting for all considerations (*clinical, statistical and operational*)

Warning when:

- <u>delayed treatment effect</u>:

futility interim → potential misleading negative early effect
superiority interim → potential lack of positive effect

- <u>early treatment effect</u>:

futility interim → potential lack of negative effect
superiority interim → potential misleading positive early effect

- *Example in a Phase III trial of tremelimumab in metastatic melanoma**:

Early interim analysis showed no survival benefit → **stop for futility**
BUT → extended follow-up showed **potential delayed separation of the survival curves** (non-significant)

**Ribas 2013: Phase III randomized clinical trial comparing tremelimumab with standard-of-care chemotherapy in patients with advanced melanoma.*

# Conclusions – trial design

- Statistical power is dependent of both magnitude and timing of the delayed treatment effect
- Very small differences in power between Log-Rank and RMSTD when modifying the non-PH parameters
- Long-term survival leads to a prolongation of trial duration
- Careful approach of interim analyses (timing, necessity)
- Potential misleading conclusions for:
  - Futility interim analysis when delayed treatment effect
  - Superiority interim analysis when early treatment effect

# References


- **Chen 2013**, “Statistical issues and challenges in immuno-oncology”, Journal for ImmunoTherapy of Cancer
- **Hoering 2017**, “End points and statistical considerations in immuno-oncology trials: impact on multiple myeloma”, Future Oncol.
- **Menis 2016**, “The European Organization for Research and Treatment of Cancer perspective on designing clinical trials with immune therapeutics”, Annals of Translational Medicine
- **Mick 2015**, “Statistical Challenges in the Design of Late-Stage Cancer Immunotherapy Studies”, Cancer Immunology Research
- **Royston 2016**,“*Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated*” BMC Medical Research methodology
- **Seymour 2017**, “iRECIST: guidelines for response criteria for use in trials testing immunotherapeutics”, Lancet Oncol
- **Sit 2016**, “Design and analysis of clinical trials in the presence of delayed treatment effect.”, Statistics in Medicine
- **Xu 2016**, “*Designing therapeutic cancer vaccine trials with delayed treatment effect*” Statistics in Medicine
- **Yang 2010**, “Improved Logrank-Type Tests for Survival Data Using Adaptive Weights”, Biometrics