Delayed treatment effects, treatment switching and heterogeneous patient populations: how to design and analyse randomized controlled trials in oncology

R. Ristl*, N. Ballarini*, H. Götte**, A. Schüler**, M. Posch* and F. König*

* Section of Medical Statistics, Medical University of Vienna
** Merck

PSI Journal Club
Online, 26 May 2022

This project has received funding from Merck

Pharmaceutical Statistics (open access): https://doi.org/10.1002/pst.2062

Received: 23 September 2019 | Revised: 16 April 2020 | Accepted: 16 July 2020
DOI: 10.1002/pst.2062

MAIN PAPER

WILEY

Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology

Robin Ristl[1] | **Nicolás M Ballarini**[1] | **Heiko Götte**[2] | **Armin Schüler**[2] | **Martin Posch**[1] | **Franz König**[1]

[1]Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria
[2]Merck Healthcare KGaA, Darmstadt, Germany

Correspondence
Franz König, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Vienna, Austria.
Email: franz.koenig@meduniwien.ac.at

Funding information
Merck Healthcare KGaA

SUMMARY
In the analysis of survival times, the logrank test and the Cox model have been established as key tools, which do not require specific distributional assumptions. Under the assumption of proportional hazards, they are efficient and their results can be interpreted unambiguously. However, delayed treatment effects, disease progression, treatment switchers or the presence of subgroups with differential treatment effects may challenge the assumption of proportional hazards. In practice, weighted logrank tests emphasizing either early, intermediate or late event times via an appropriate weighting function may be used to accommodate for an expected pattern of non-proportionality. We model these sources of non-proportional hazards via a mixture of survival functions with piecewise constant hazard. The model is then applied to study the power of unweighted and weighted log-rank tests, as well as maximum tests allowing different time dependent weights. Simulation results suggest a robust performance of maximum tests across different scenarios, with little loss in power compared to the most powerful among the considered weighting

Survival as primary endpoint

- Survival time is the most relevant outcome in oncology trials
- We focus on randomized controlled trials with two groups (treatment versus control)
- Aim: Compare distributions of survival times T to show superiority of treatment
- Problem: Survival times are censored for patients who are alive at final follow up
- Cannot compare, e.g., means without parametric model
- Standard approach: compare hazard functions

Hazard at time t: $\lambda(t) = \lim_{h \downarrow 0} P(T \in (t, t+h] | T > t)/h$

Survival distribution: $S(t) = P(T > t) = \exp\left\{-\int_0^t \lambda_i(s)ds\right\}$

(Weighted) Logrank test

- Standard test for censored data to compare two survival distributions via their hazard functions.
- Null hypothesis: $H_0 : \lambda_{trt}(t) \geq \lambda_{ctr}(t), \forall t \geq 0$
- Test statistic

$$z = \sum_{t \in \mathcal{D}} w(t)(d_{t,ctr} - e_{t,ctr}) / \sqrt{\sum_{t \in \mathcal{D}} w(t)^2 var(d_{t,ctr})}$$

d number of observed events in control group at time t
e number of expected events in control group under least favorable configuration in H_0.
$\mathcal{D}$ is the set of all observed event times
$w(t)$ weight for contribution at time t

- Time stratified Cochran-Mantel-Haenszel test
- Asymptotically z is standard normal under least favorable configuration in H_0.

Proportional hazards (PH) assumption

$$\lambda_{ctr}(t)/\lambda_{trt}(t) = const.$$

Sample size planning and interpretation of logrank test are typically made under this assumption. It is also the underlying assumption of Cox regression. Under PH

- Logrank test is the most powerful rank invariant test for H_0 (Peto 1972)
- Rejection of H_0 implies $P_{trt}(T > t) > P_{ctr}(T > t), \forall t$
- Simple asymptotic relations for sample size calculation

But:

- PH assumption questionable for many oncology studies
- In particular, new generation of immune-oncology drugs require time to unfold efficacy, resulting in observation of so called "delayed onset" or "late separation" of survival functions (Anagnostou 2017).
- Starting point for this work

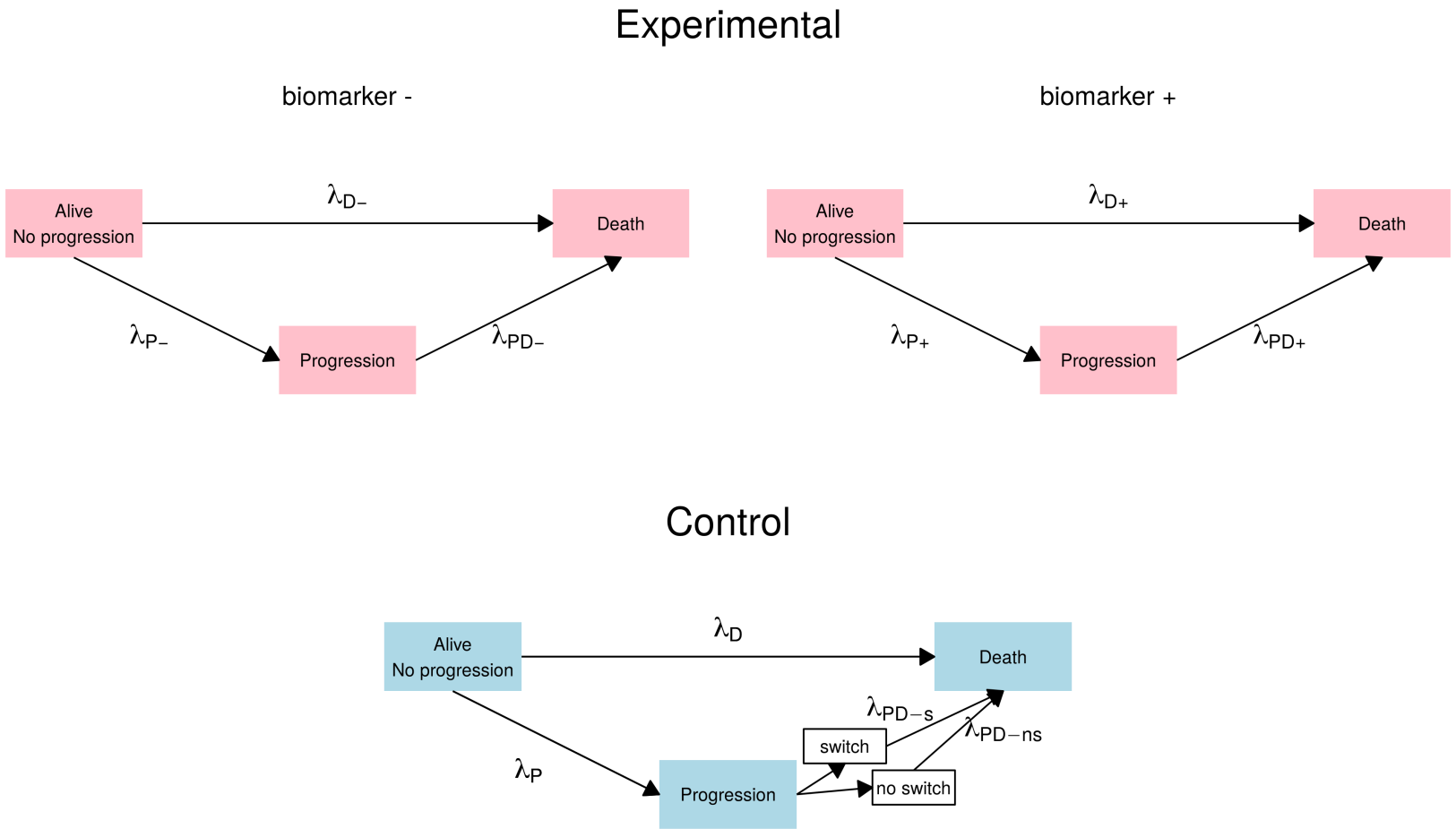
Aims

- Identify relevant scenarios of non-proportional hazards
- Propose simple but flexible model for survival functions under non-proportional hazards
- Compare power of different hypothesis tests under the identified scenarios
- Perform conditional power calculations at interim analysis

Sources of non-proportional hazards

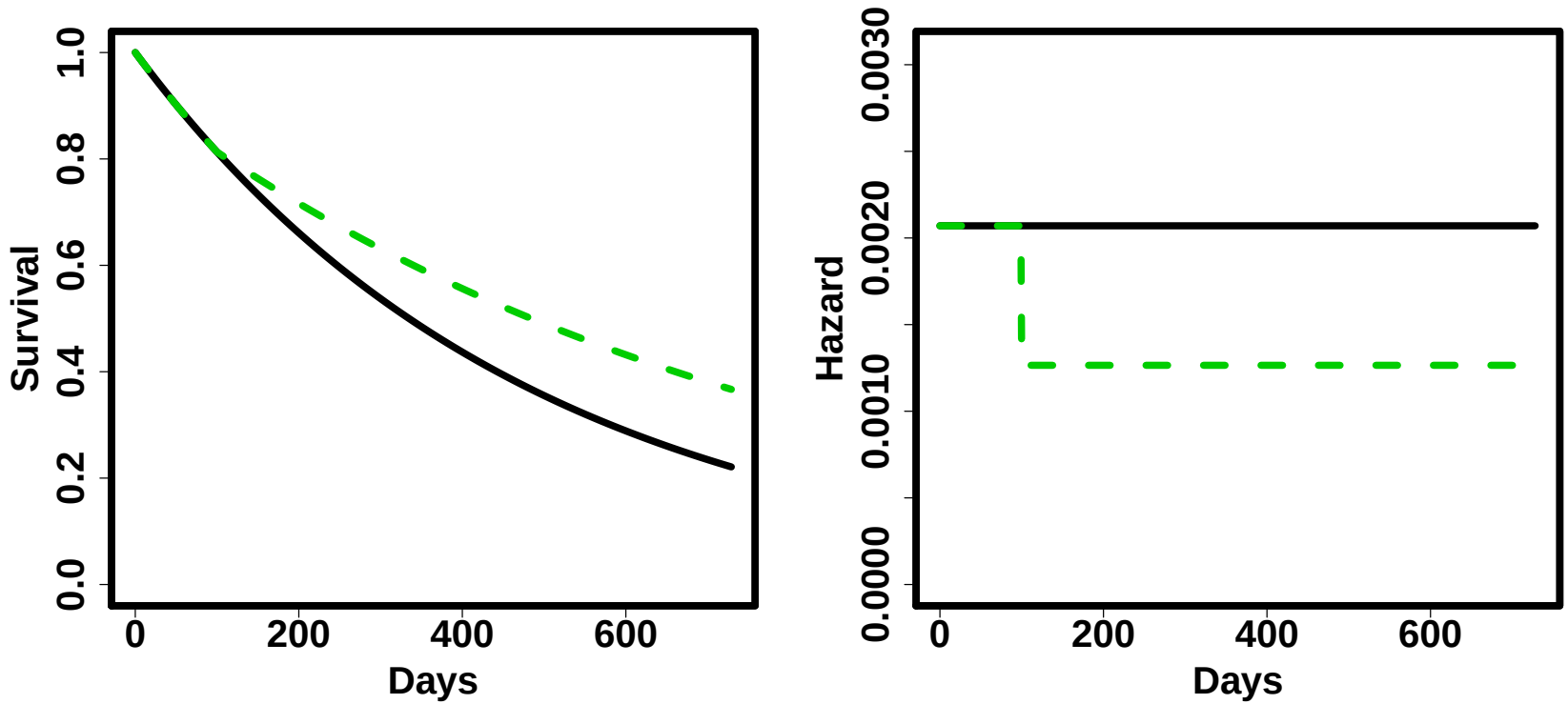
- Delayed onset of treatment effect
  - No treatment effect at early event times
- Changing hazards after disease progression
  - Progression rate and subsequent hazards may depend on treatment
- Biomarker subgroups
  - Patients positive for a specific biomarker may show increased treatment benefit (e.g. EGFR inhibitors are effective only against tumours that are free from mutations in the KRAS or NRAS genes, Chan 2017)
  - Composition of treatment group study population changes with time, as biomarker positive patients survive longer
- Treatment switching after disease progression
  - Control group patients may switch to treatment group medication
  - Observed treatment effect at late event times is reduced

Proposed model



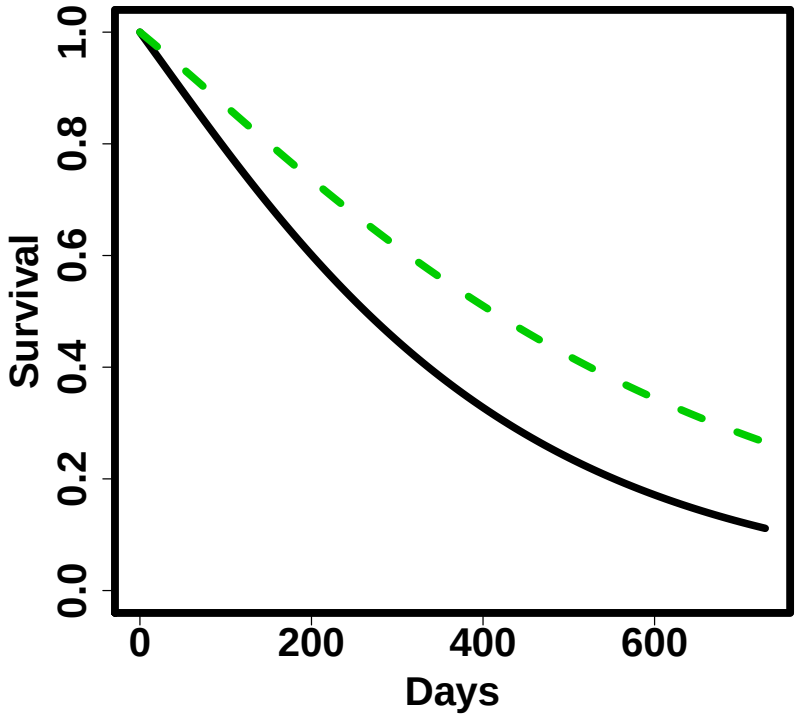
States are represented by boxes. Hazard functions $\lambda_.$ for transitions are indicated next to the respective arrows. All hazard functions are modelled as piecewise constant functions of time.

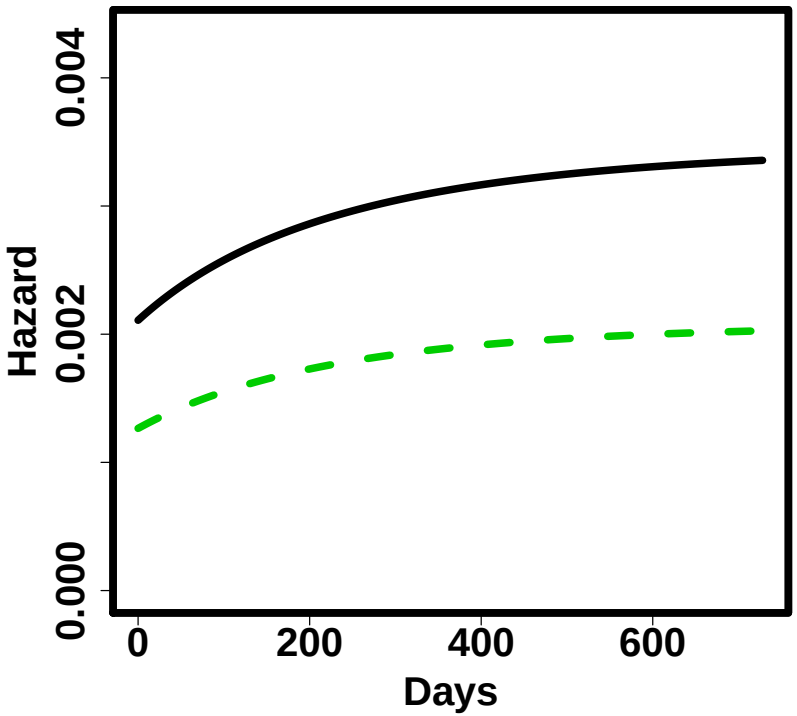
Example for model with delayed onset



Treatment effect starts after 100 days. Constant hazard rate under control (black) and treatment (green) up to 100 days corresponding to median survival of 11 months. Constant hazard rate under treatment after 100 days corresponding to median of 18 months.

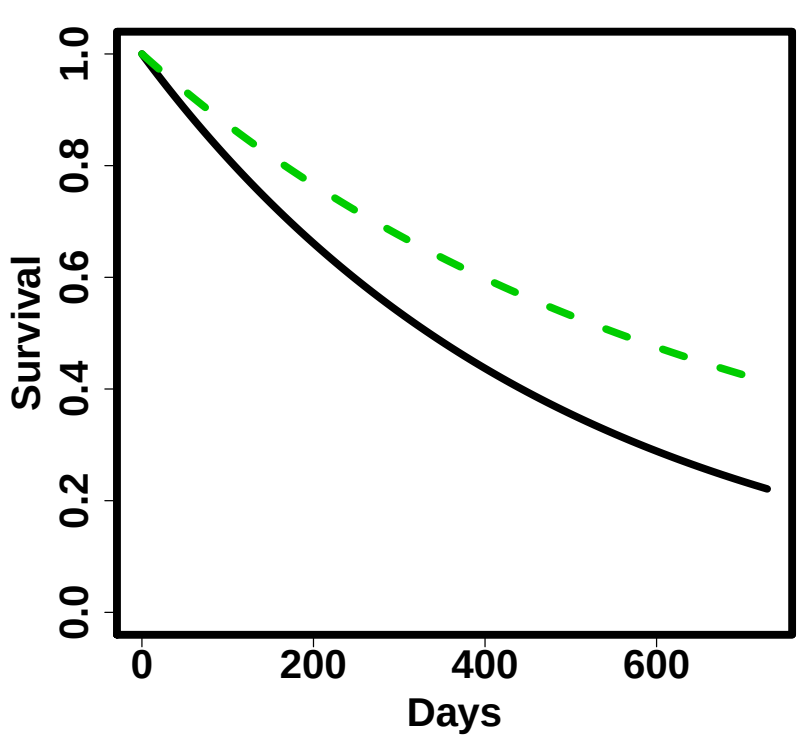
Changing hazards after disease progression

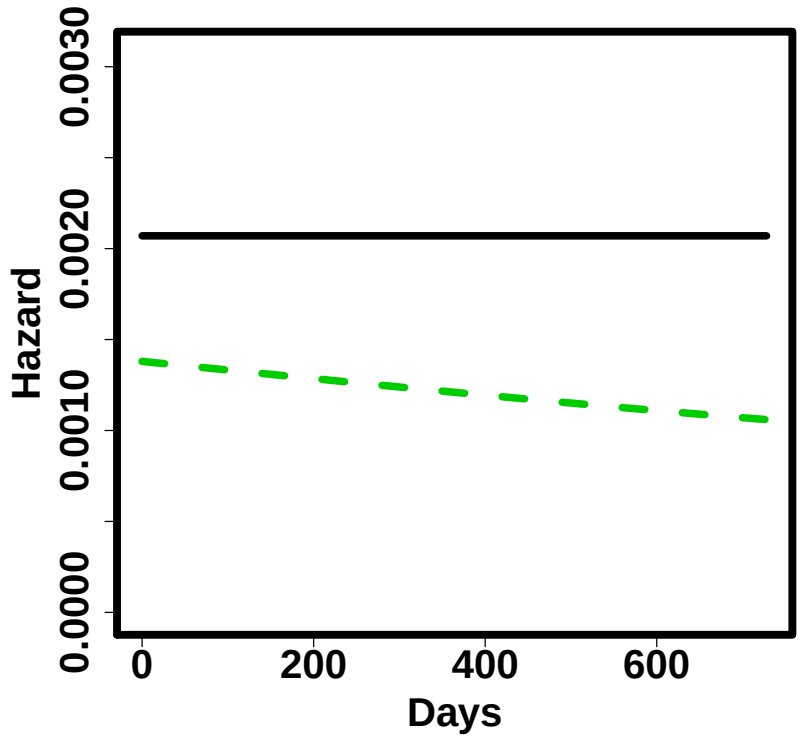




Hazard rate under control corresponds to median survival of 18 months before and 11 months after progression. Hazard ratio of 0.6 between treatment and control for both rates. Common rate of progression with median progression time of 5 months.

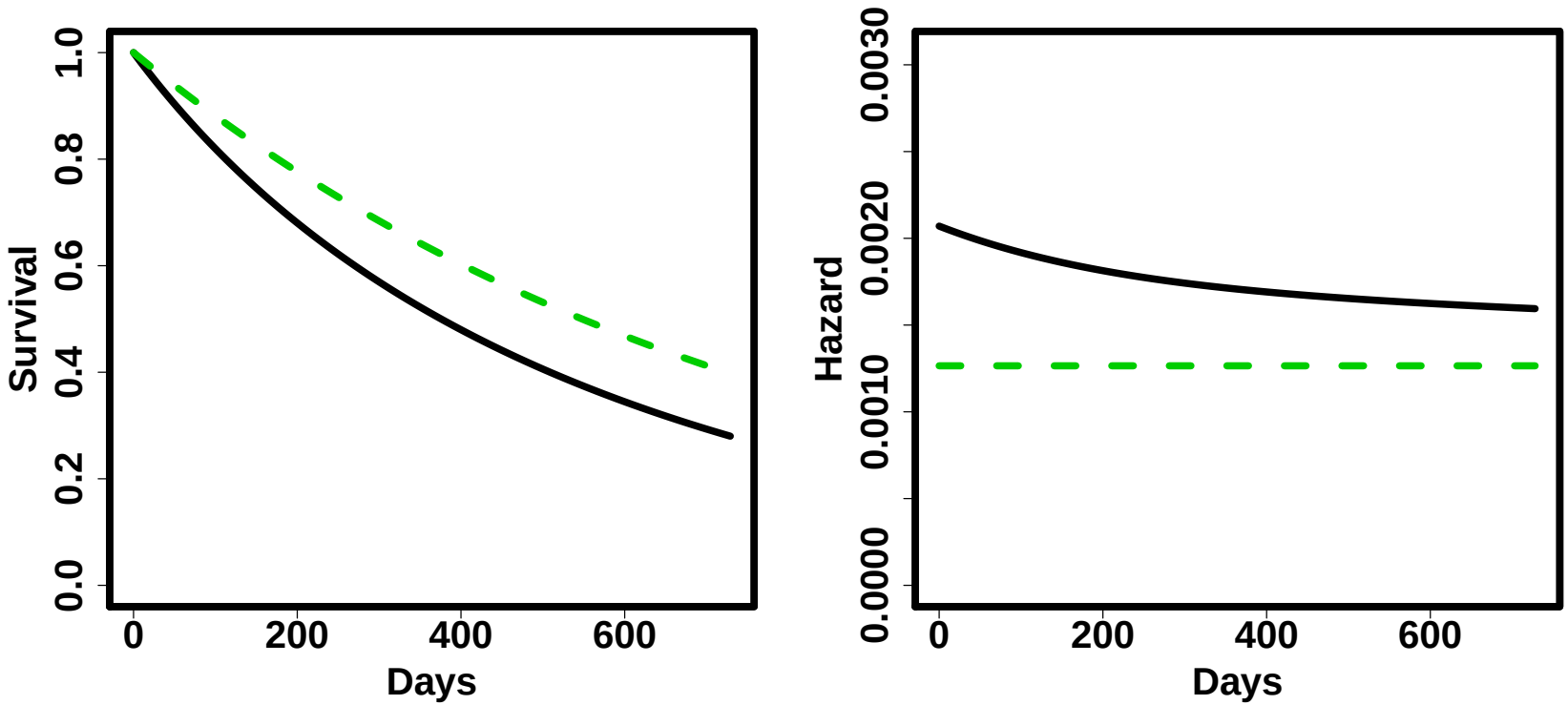
Biomarker subgroups





50% of patients are biomarker positive, their median survival is 33 months versus 11 months for biomarker negative and control patients. (Constant hazards, but composition of population changes with time.)

Treatment switching



After progression, control patients switch to treatment with 50% probability. Constant hazard rates corresponding to median survival of 18 (treatment) versus 11 (control) months. Median time to progression is 5 months in both groups.

Hypothesis testing

We consider weighted logrank tests to put more weight on early, intermediate or late event times.

- Standard logrank test: $w(t) = 1$
- Fleming-Harrington $\rho - \gamma$ family: $w(t) = \hat{S}(t)^{\rho}(1 - \hat{S}(t))^{\gamma}$. We consider $(\rho, \gamma) \in \{(0,1), (1,1), (1,0)\}$
- Maximum test:
  - k weight functions $w_1(t), \ldots, w_k(t)$ (from $\rho - \gamma$ family)
  - Corresponding weighted log-rank statistics $z_1, \ldots, z_k$
  - Maximum test statistic $z_{max} = \max_{i=1,\ldots,k} |z_i|$
  - Under H_0, approximately $Z_{max} \sim N_k(0, \Sigma)$
  - To estimate Σ use $cov(w_i(t)d_{t,ctr}, w_j(t)d_{t,ctr}) = w_i(t)w_j(t)var(d_{t,ctr})$ (assume weights non-random or converging in probability to non-random function)
  - Calculate p-value $P_{H_0}(Z_{max} > z_{max})$ from multivariate normal distribution
  - See Tarone 1981, approach has recetly received some attention.
- Alternative: Modestly weighted logrank tests Magirr and Burman, 2018

Different emphasis of Fleming-Harrington weights

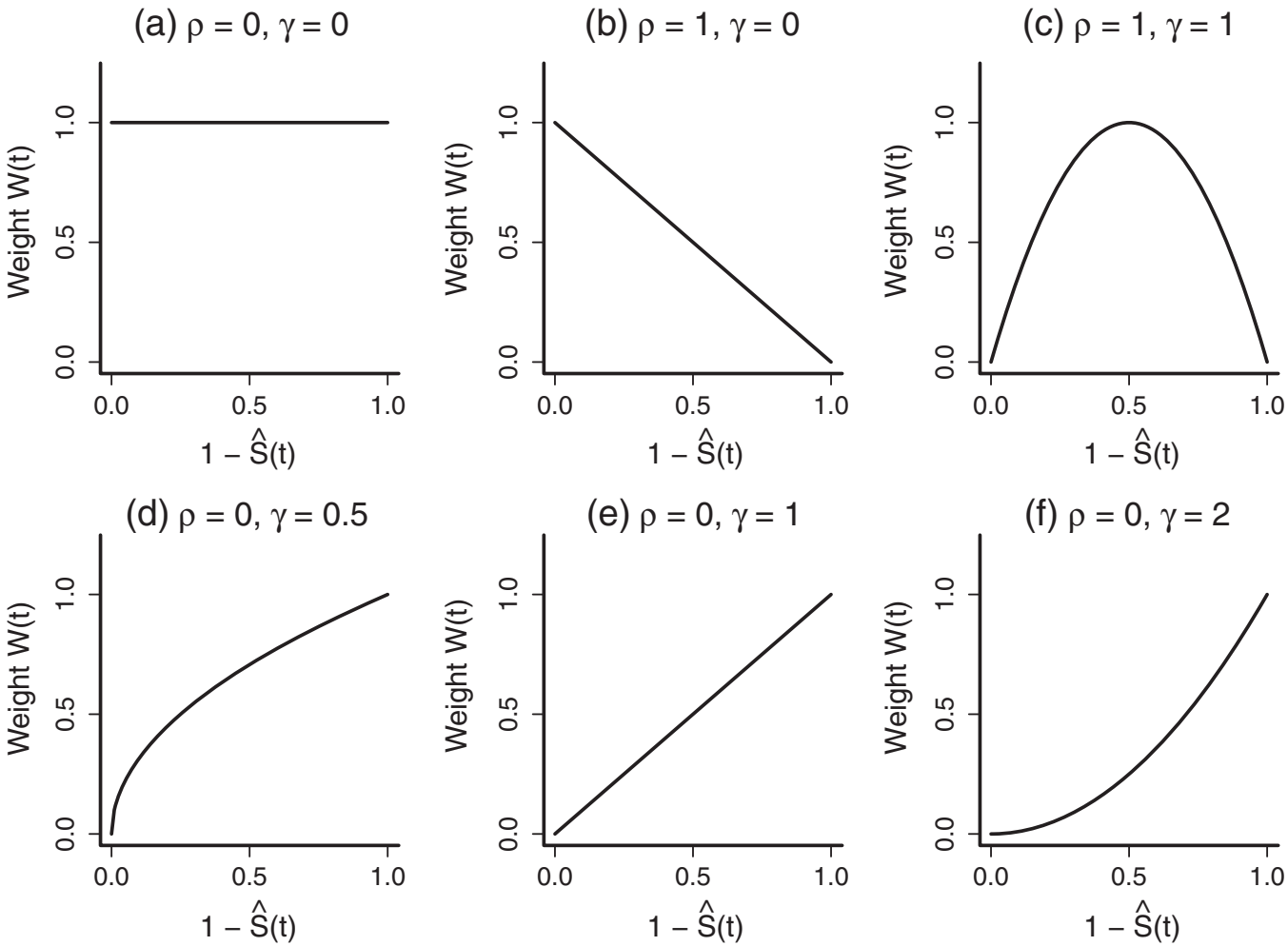


Figure 1 from T. Hasegawa, Pharmaceut. Statist. 2014, 13 128-135

Simulation scenario for randomized controlled trial

- Equal allocation ratio to treatment and control
- 1 year recruitment at rate of 300 patients/year
- Analysis after 130 events in total have been observed (appr. 1.5 years with assumed scenarios)
  - Under proportional hazards this number of events would correspond to approximately 80% power of the logrank test with a hazard ratio of 0.6 at one-sided level of significance of 0.025.
- No random censoring due to drop outs
- Survival times drawn from piecewise constant hazards models shown before
- 10,000 simulation runs per scenario

Power (%) under the four scenarios

Test weight (ρ, γ)	Delayed onset	Progression effect	Biomarker Subgroups	Treatment switching
Equal $(0, 0)$	35	81	75	60
Early $(1, 0)$	27	80	72	61
Intermediate $(1, 1)$	49	73	70	46
Late $(0, 1)$	49	68	66	40
Maximum all	45	79	73	56
Maximum $(0, 0)$ $(0, 1)$	45	79	73	55

- Best test depends on scenario.
- Maximum test typically has power close to best included test.
- Standard logrank test performs well in many non-proportional hazard scenarios.
- Similar observations were made under more complex scenarios.

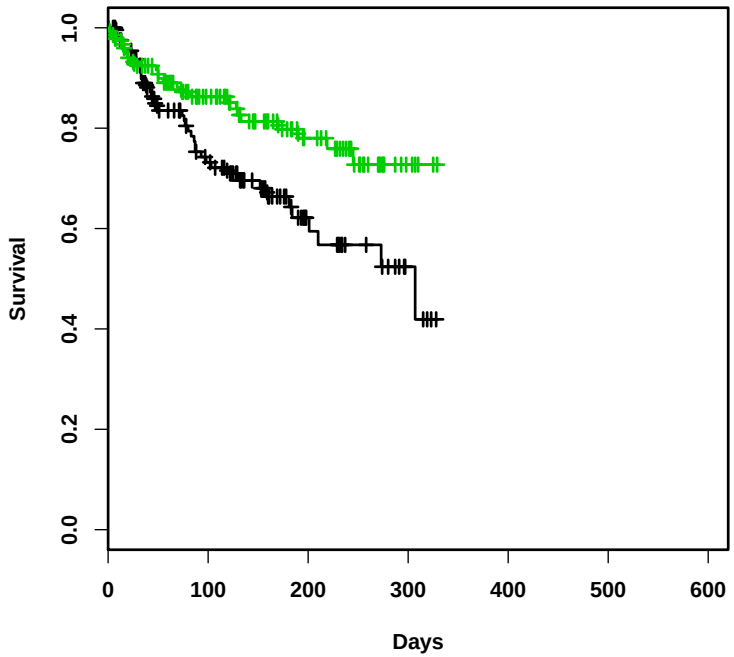
Conditional power based on interim data

- Assume interim data after certain number of events
- Want to calculate conditional power for final analysis after planned number of events
- Under design assumptions, marginal survival function is known
- For a given censored observation y_{cens}, we may calculate the conditional survival function $P(Y > t | Y > y_{cens})$
- Sample survival times for observations censored at interim from conditional distribution
- Also sample additional patients if recruitment ongoing
- Thus, sample data set with planned number of events, conditional on interim data
- Use multiple conditional samples to calculate conditional power
- Stopping for futility based on conditional power

Conditional data simulation example

True: Delayed onset of treatment action after 100 days.
Hazard rates (per year): $\lambda_{ctr} = \lambda_{trt,<100d} = 0.002, \lambda_{trt,\geq 100d} = 0.001$
Recruitment 1 years, with rate 300 patients/year
Events at interim: 65, events at study end: 130

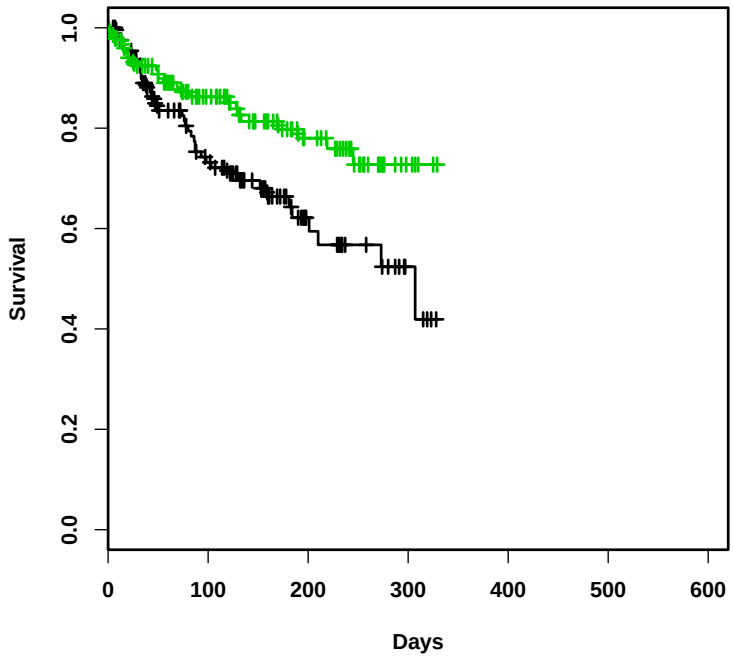
Observed Interim Data



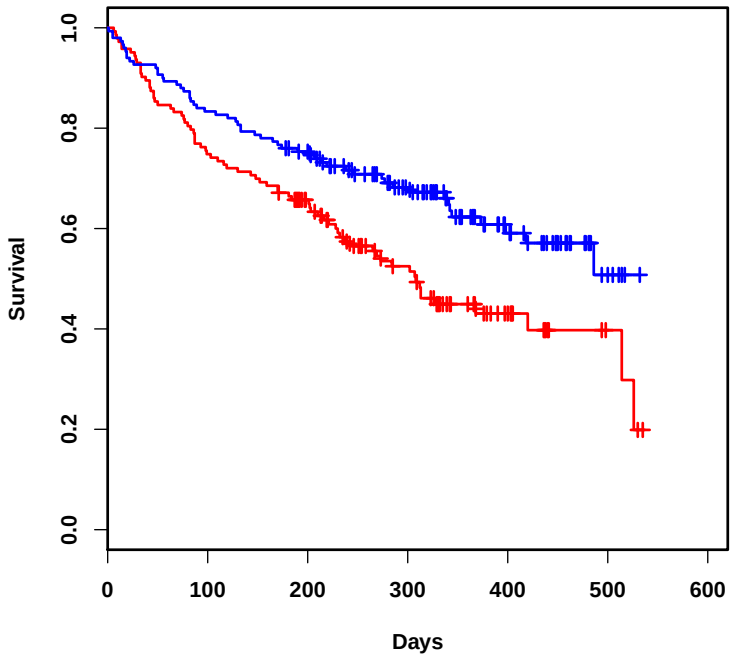
Conditional data simulation example

True: Delayed onset of treatment action after 100 days.
Hazard rates (per year): $\lambda_{ctr} = \lambda_{trt,<100d} = 0.002, \lambda_{trt,\geq 100d} = 0.001$
Recruitment 1 years, with rate 300 patients/year
Events at interim: 65, events at study end: 130

Observed Interim Data



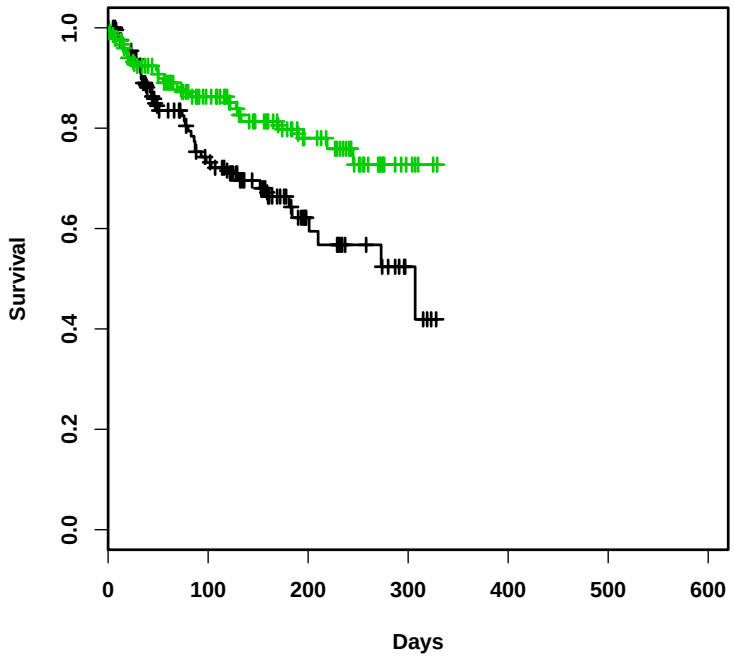
Simulated Delayed



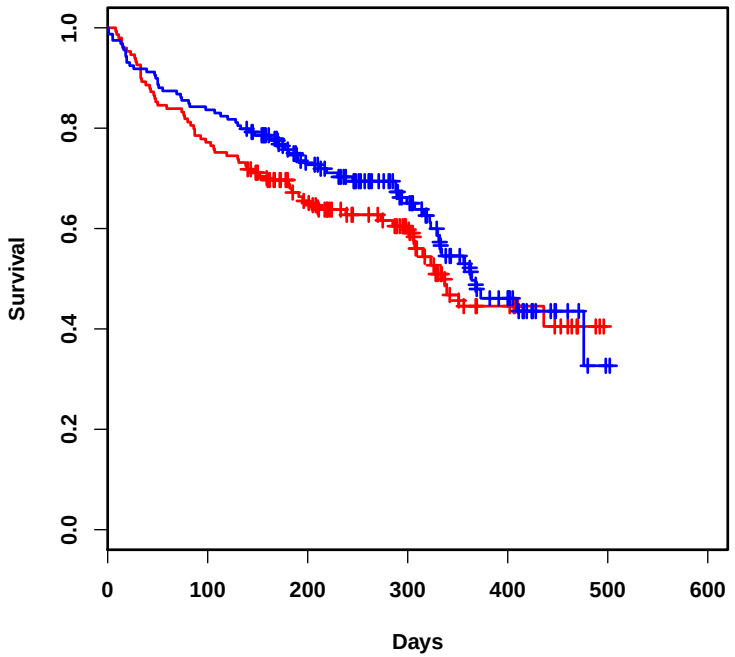
Conditional data simulation example

True: Delayed onset of treatment action after 100 days.
Hazard rates (per year): $\lambda_{ctr} = \lambda_{trt,<100d} = 0.002, \lambda_{trt,\geq 100d} = 0.001$
Recruitment 1 years, with rate 300 patients/year
Events at interim: 65, events at study end: 130

Observed Interim Data



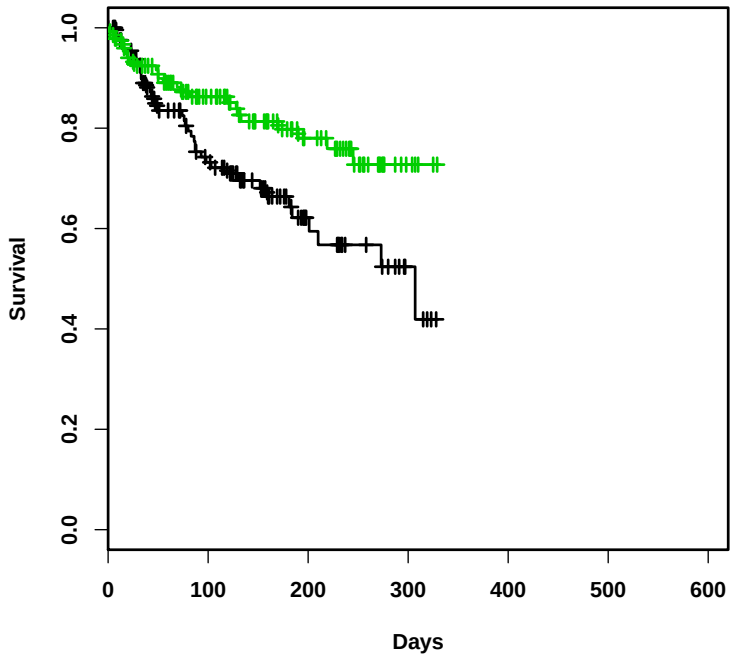
Simulated Delayed



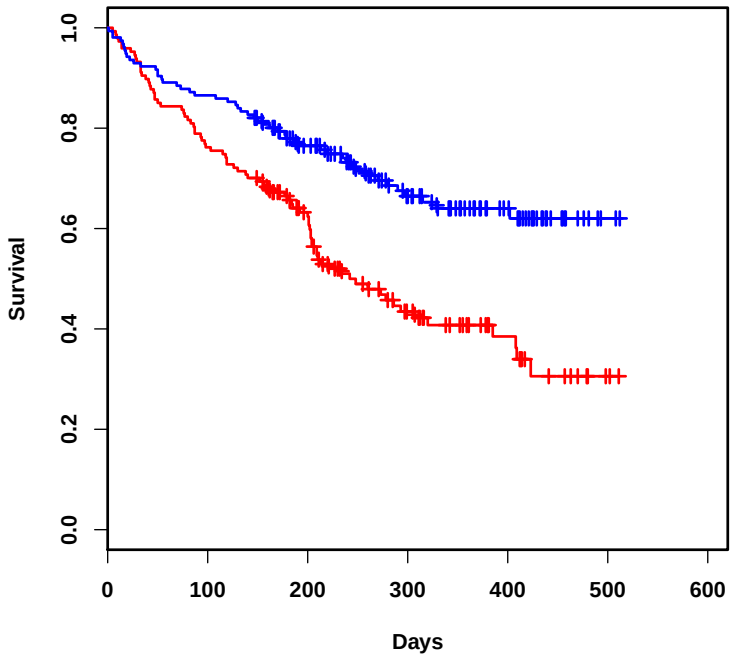
Conditional data simulation example

True: Delayed onset of treatment action after 100 days.
Hazard rates (per year): $\lambda_{ctr} = \lambda_{trt,<100d} = 0.002, \lambda_{trt,\geq 100d} = 0.001$
Recruitment 1 years, with rate 300 patients/year
Events at interim: 65, events at study end: 130

Observed Interim Data



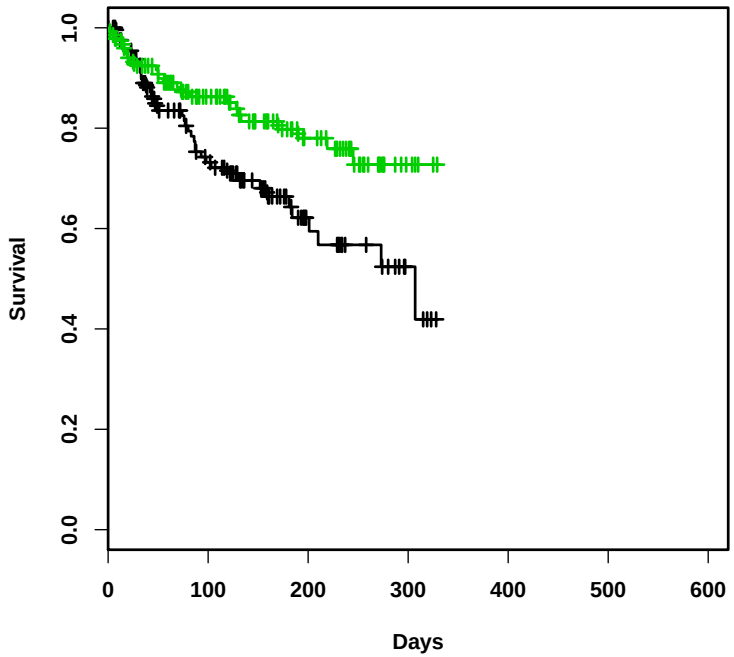
Simulated Delayed



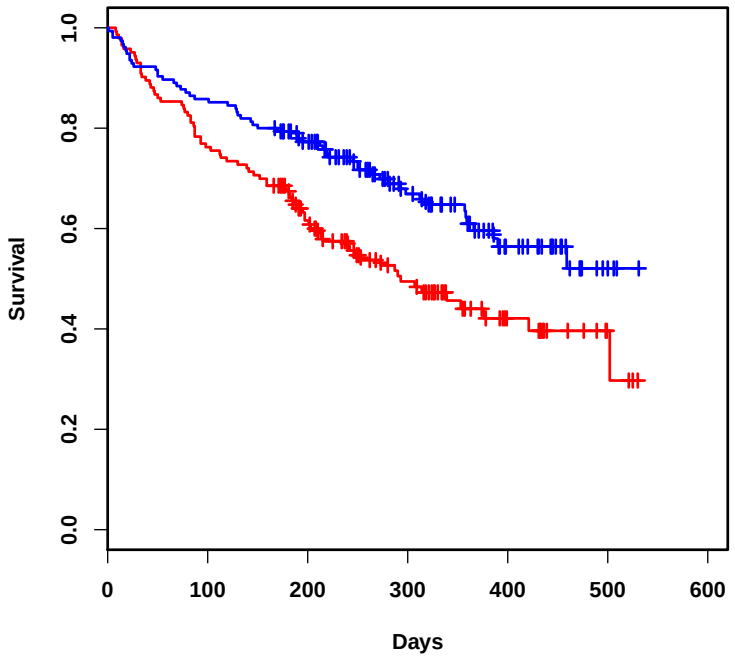
Conditional data simulation example

True: Delayed onset of treatment action after 100 days.
Hazard rates (per year): $\lambda_{ctr} = \lambda_{trt,<100d} = 0.002, \lambda_{trt,\geq 100d} = 0.001$
Recruitment 1 years, with rate 300 patients/year
Events at interim: 65, events at study end: 130

Observed Interim Data



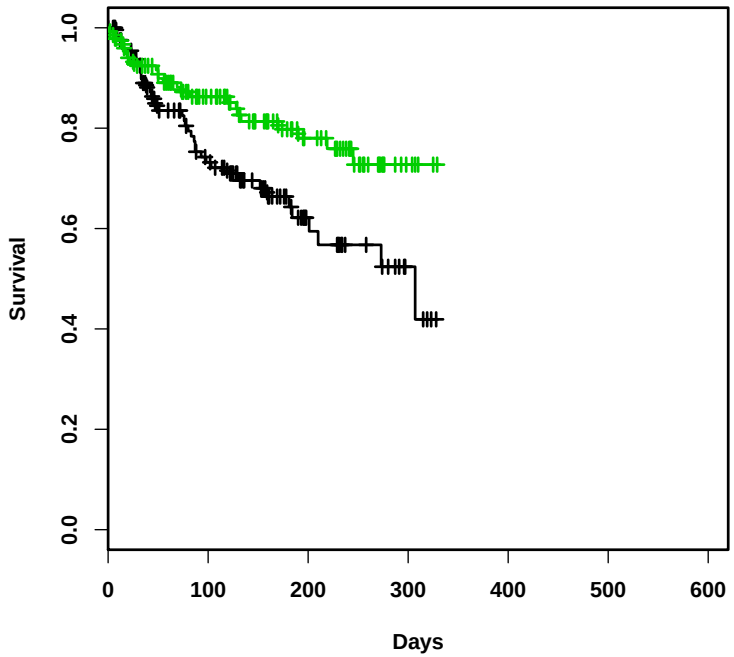
Simulated Delayed



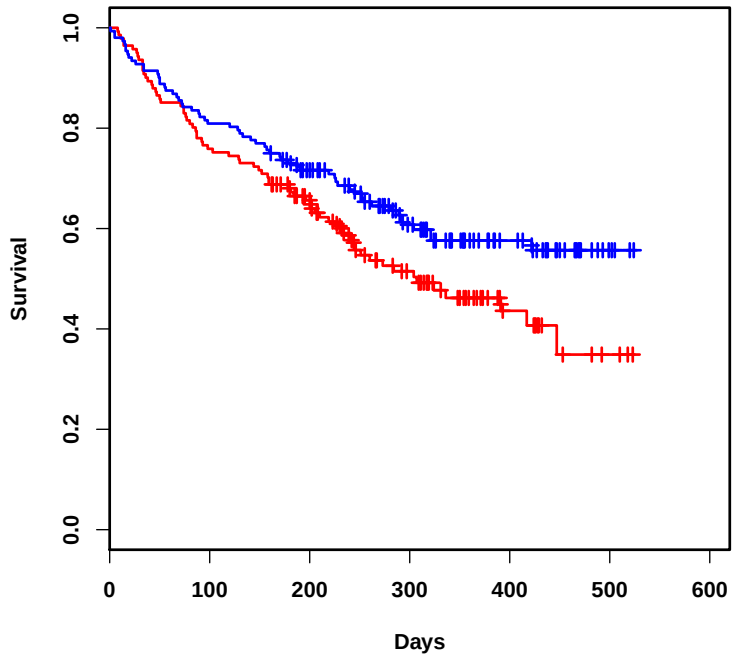
Conditional data simulation example

True: Delayed onset of treatment action after 100 days.
Hazard rates (per year): $\lambda_{ctr} = \lambda_{trt,<100d} = 0.002, \lambda_{trt,\geq 100d} = 0.001$
Recruitment 1 years, with rate 300 patients/year
Events at interim: 65, events at study end: 130

Observed Interim Data

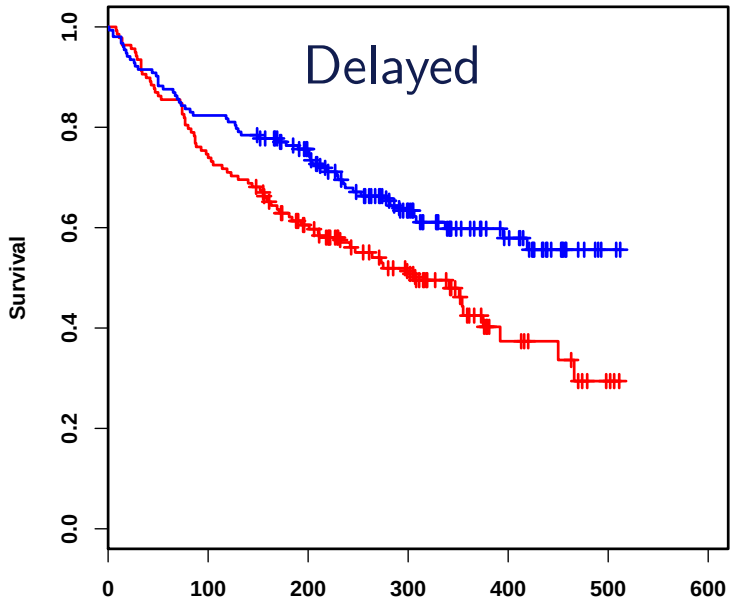


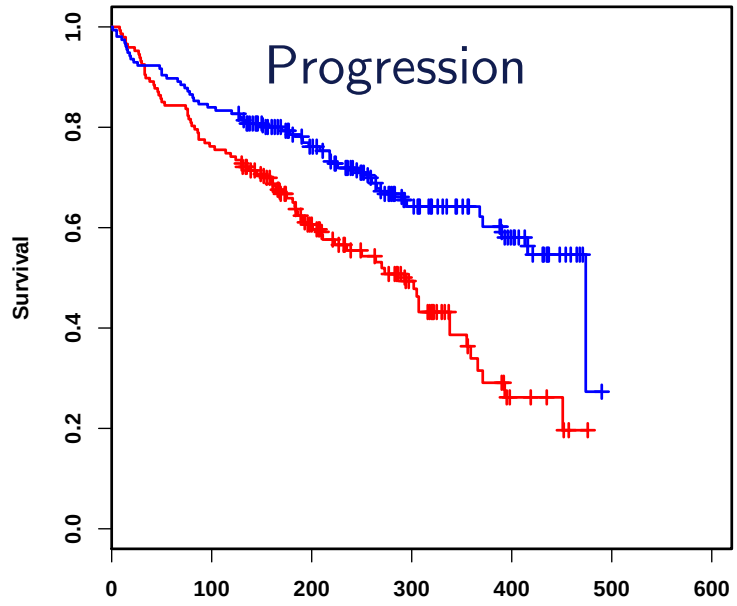
Simulated Delayed

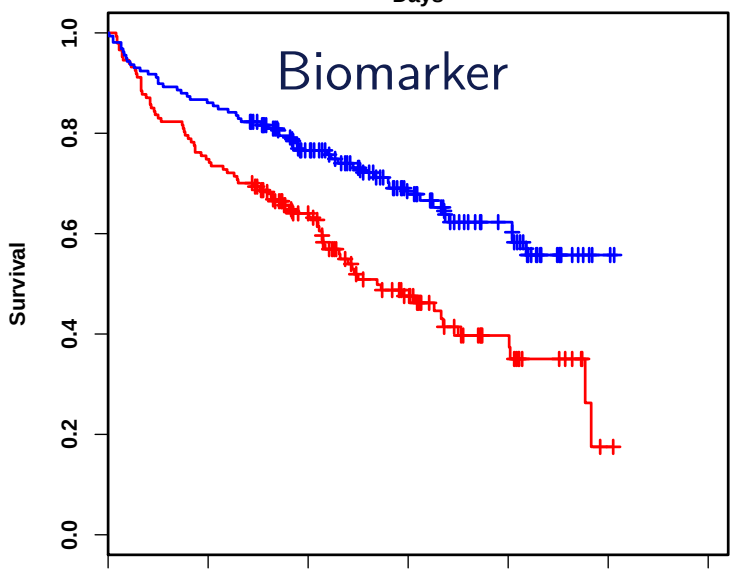


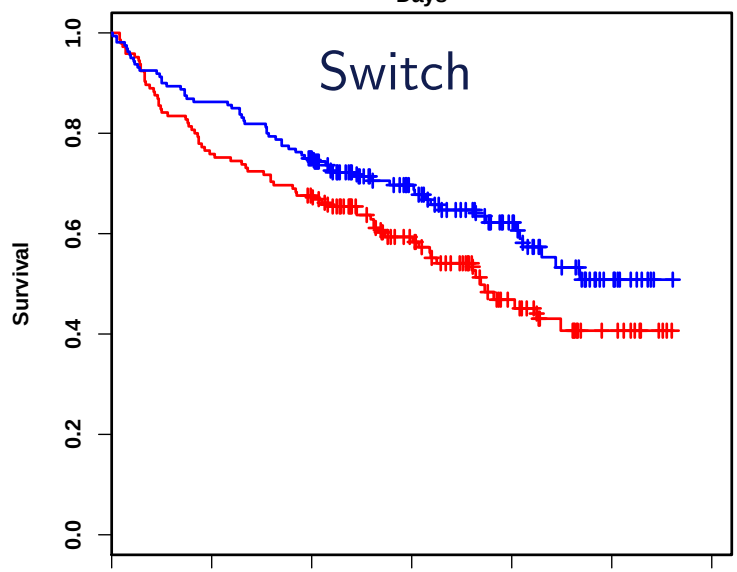
Conditional power Logrank test 92%, Maximum test 89%

Conditional Power assuming different scenarious

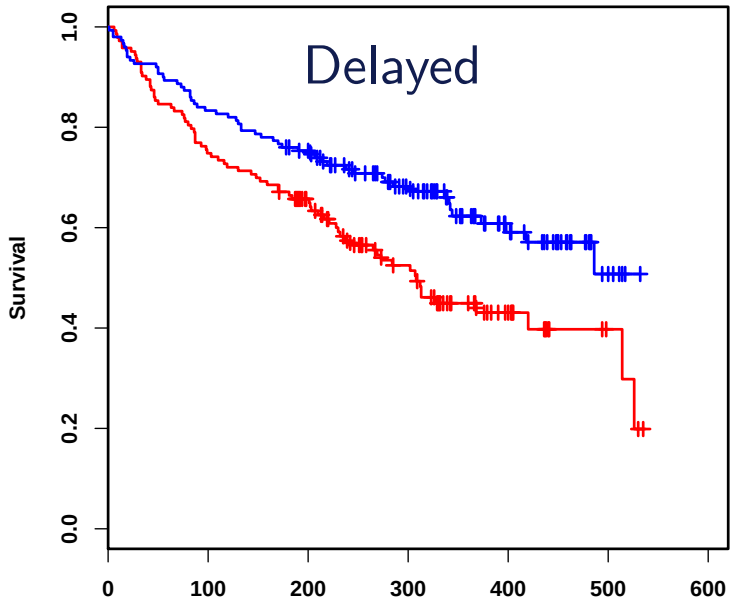


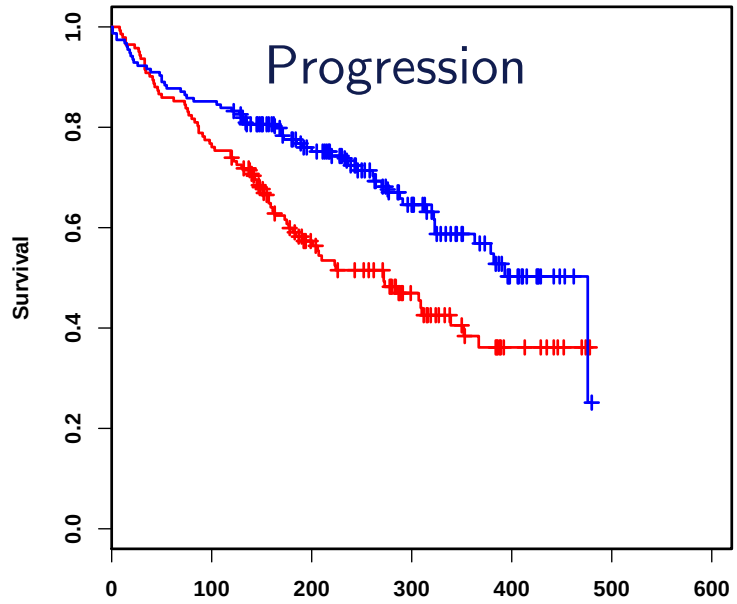


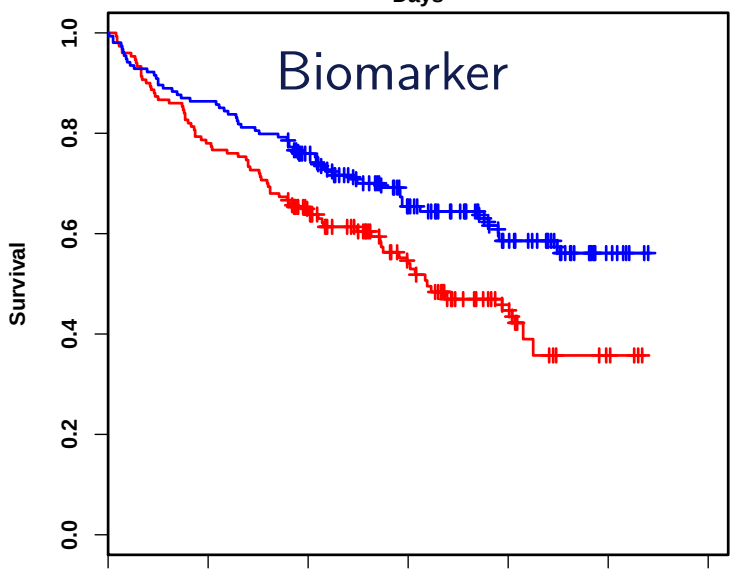


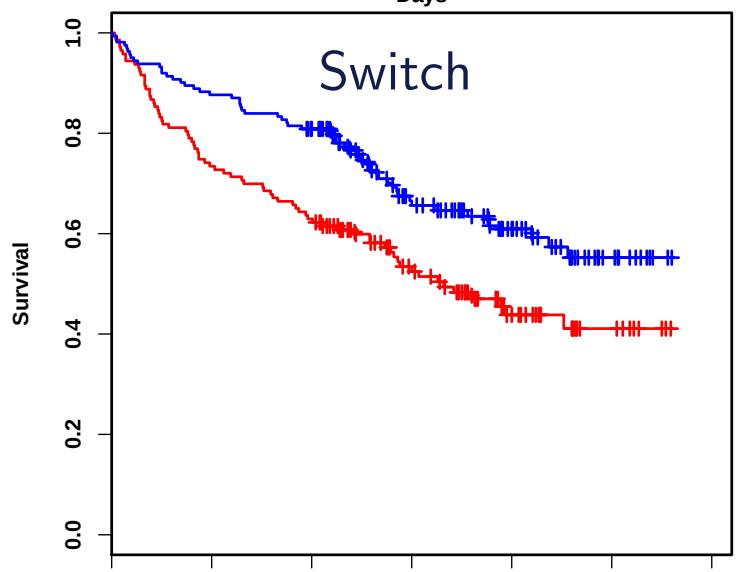


Conditional Power assuming different scenarious

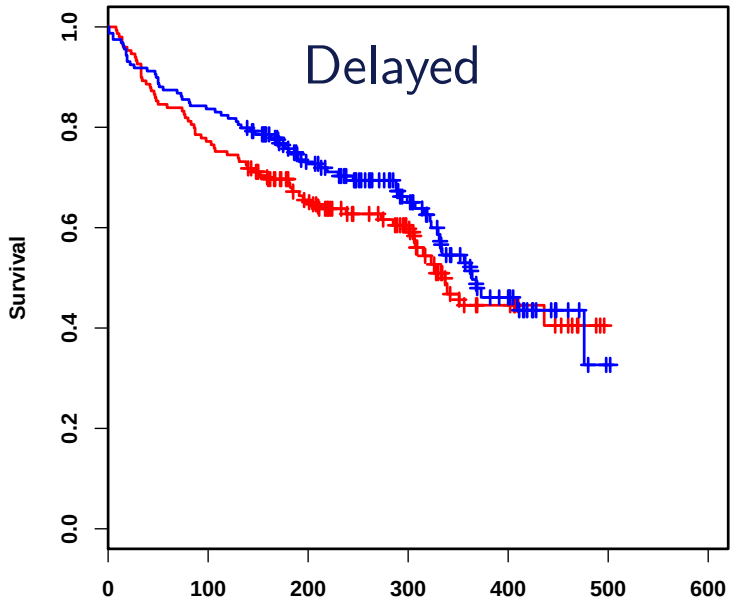


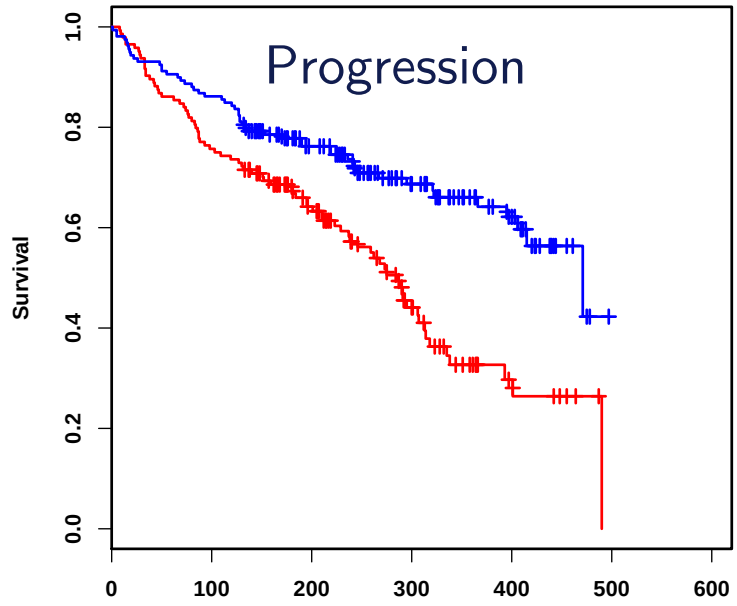


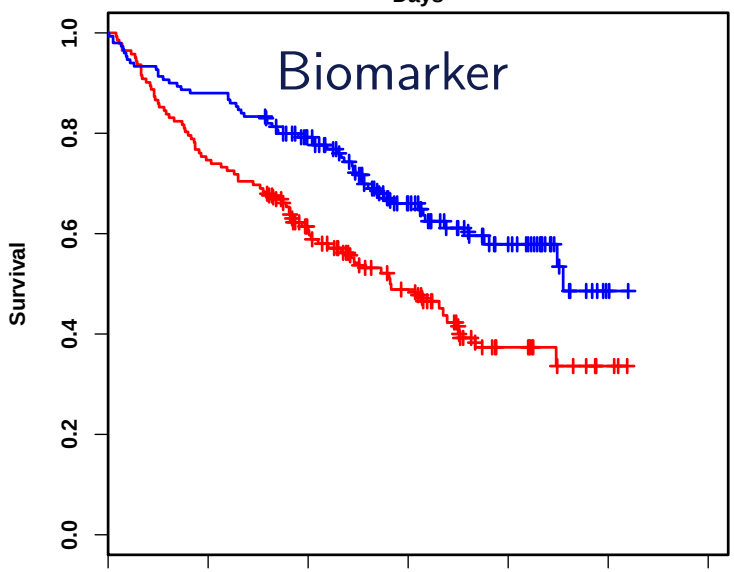


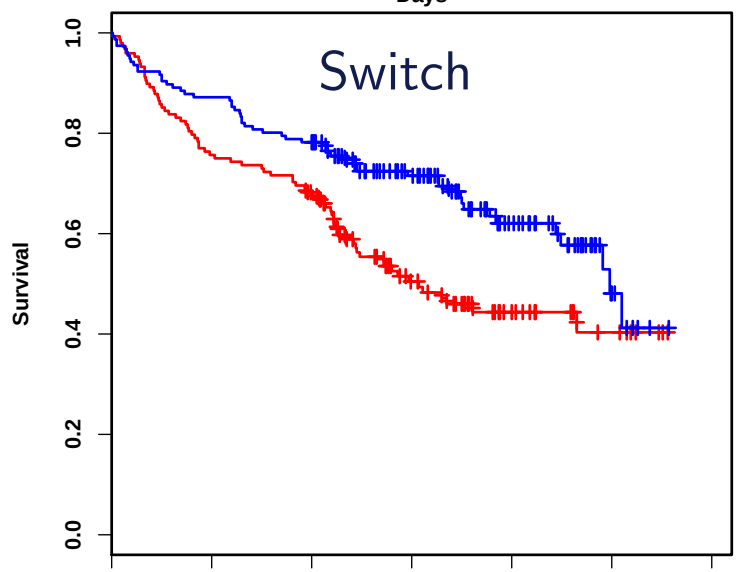


Conditional Power assuming different scenarious

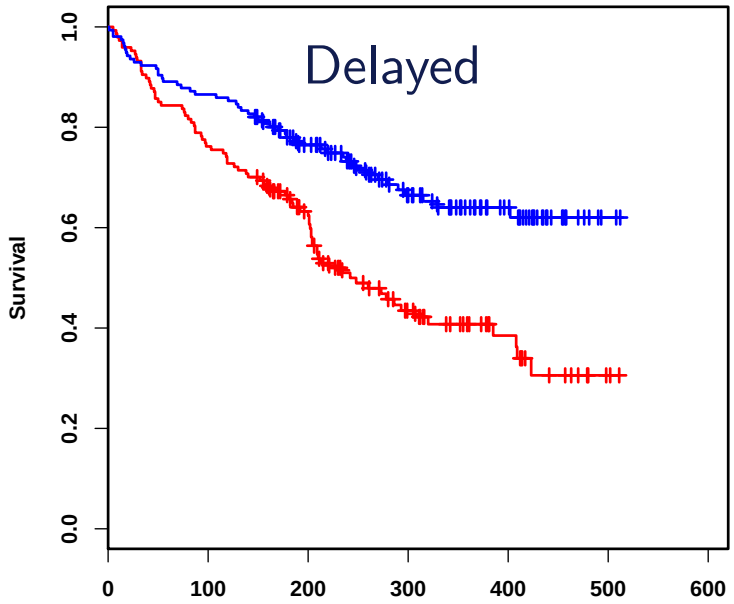


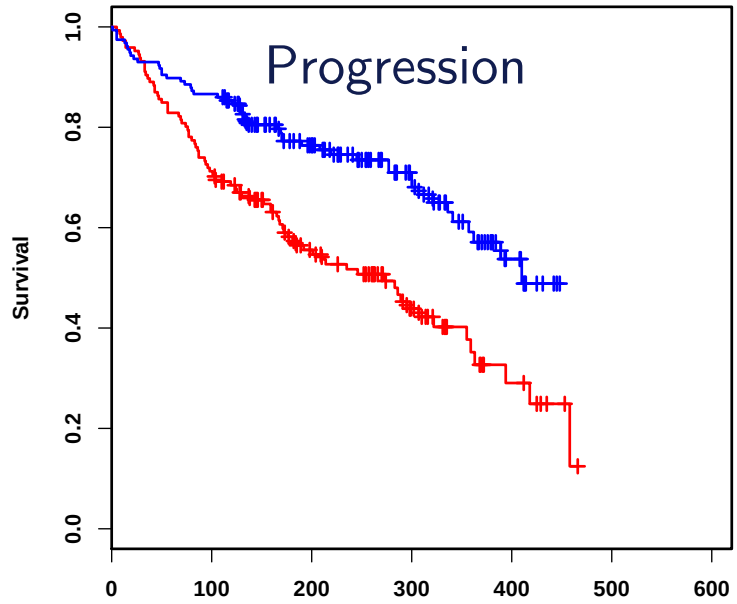


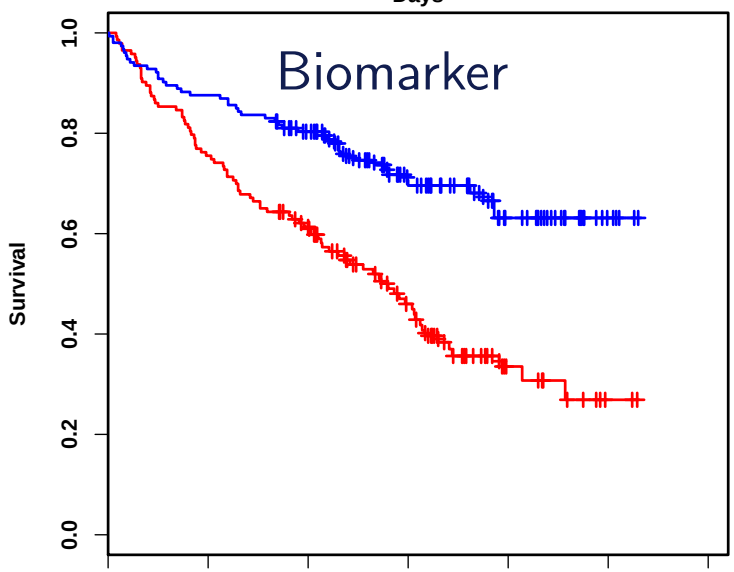


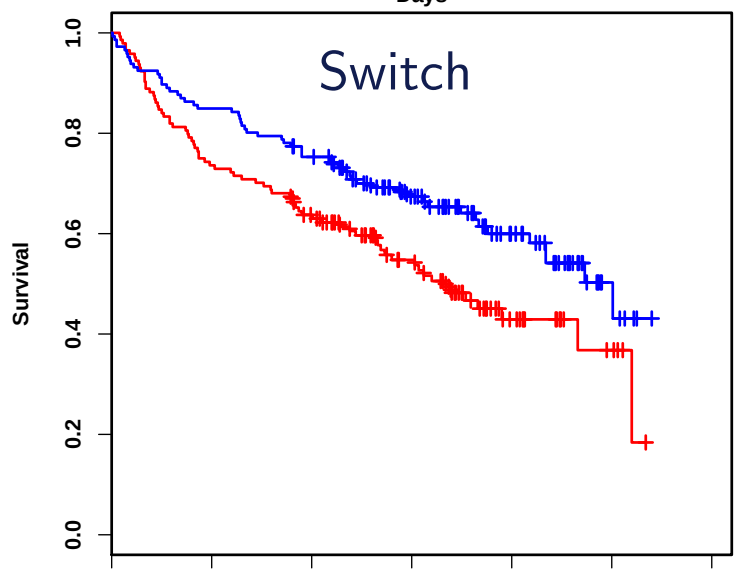


Conditional Power assuming different scenarious

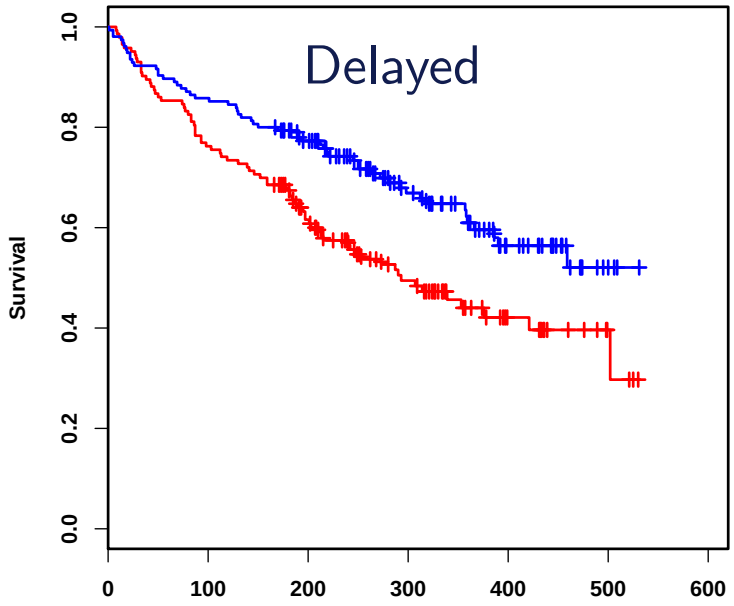


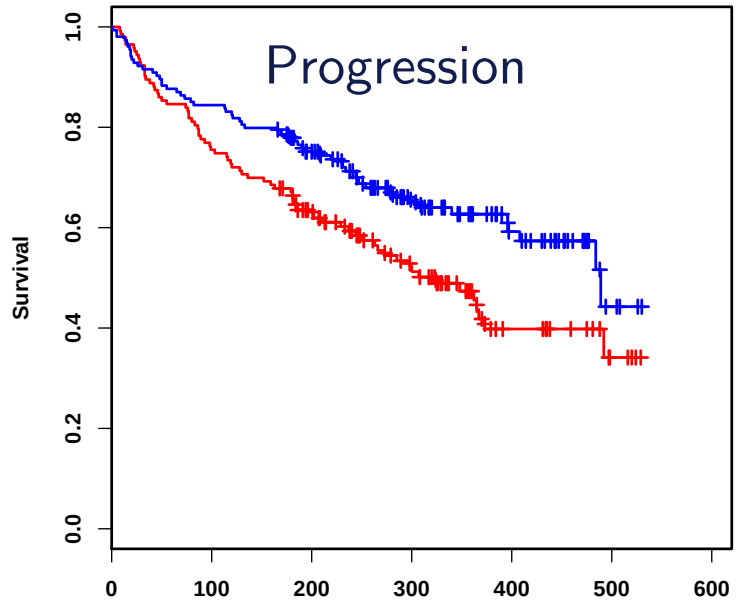


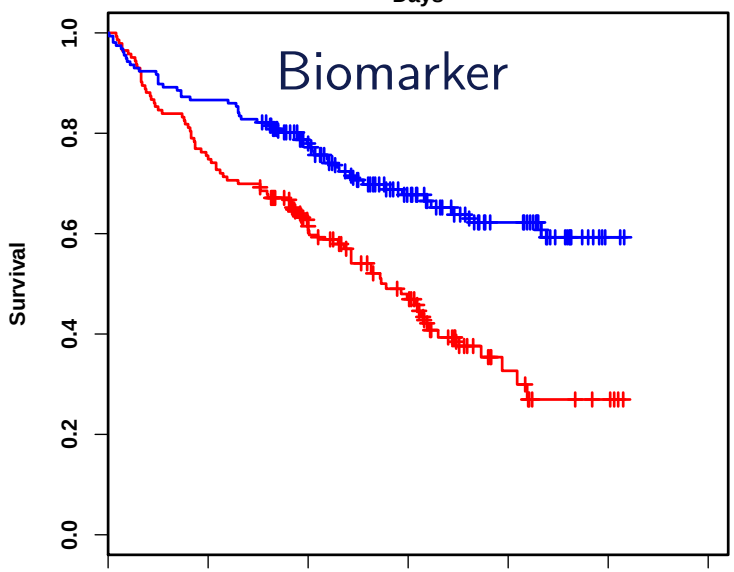


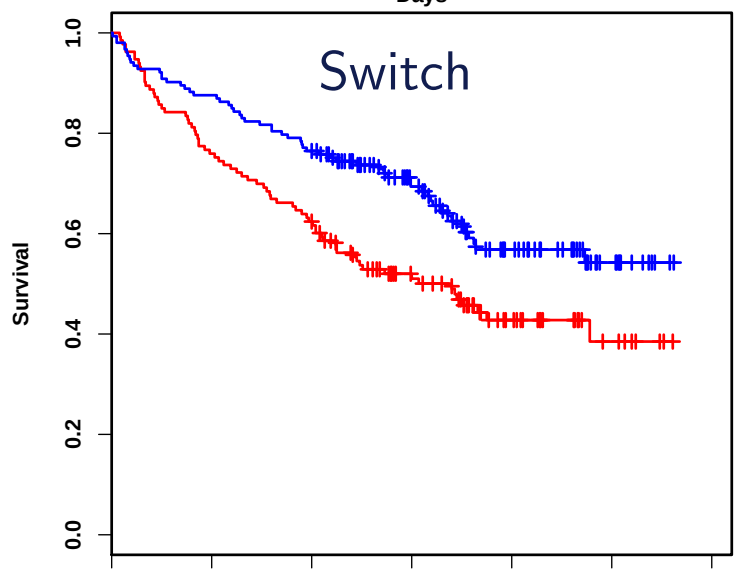


Conditional Power assuming different scenarious

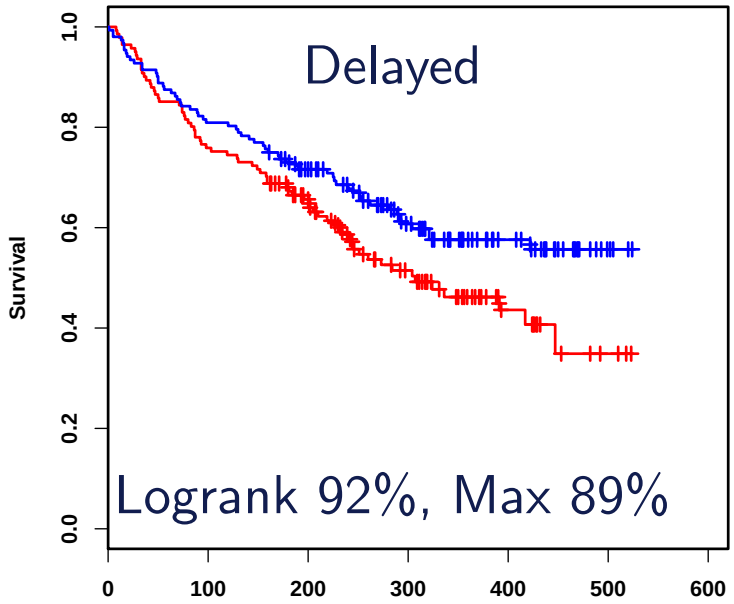


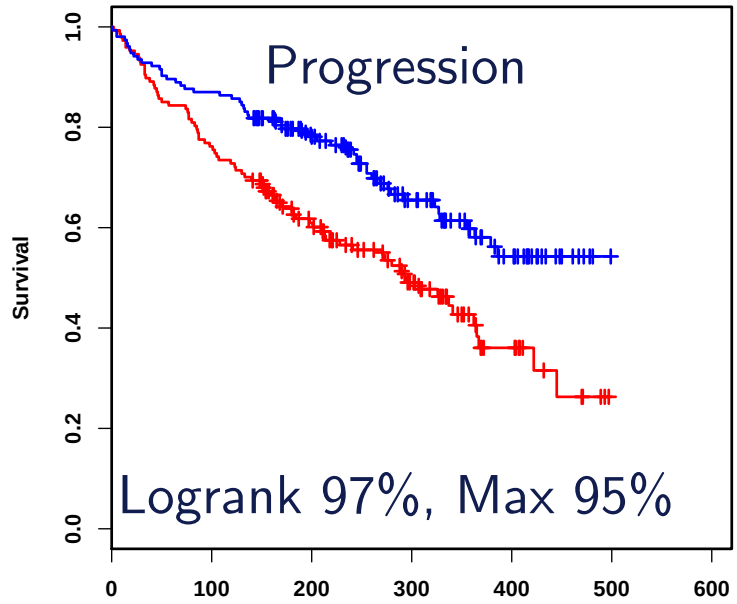


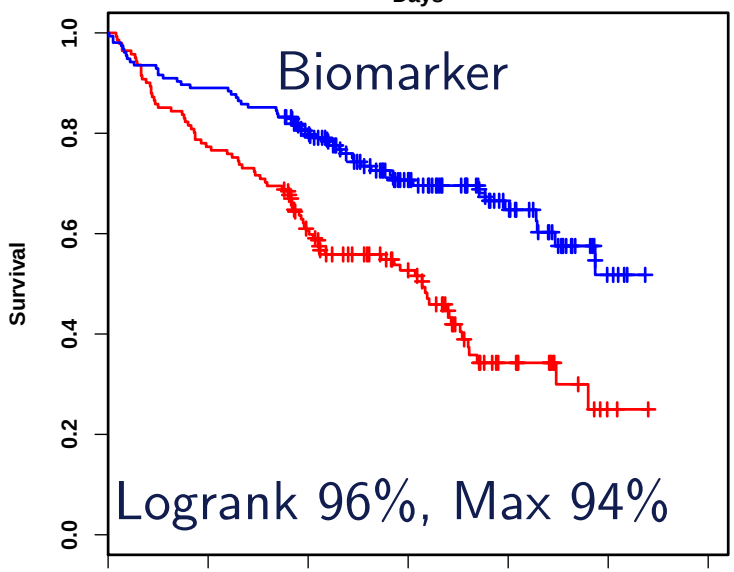


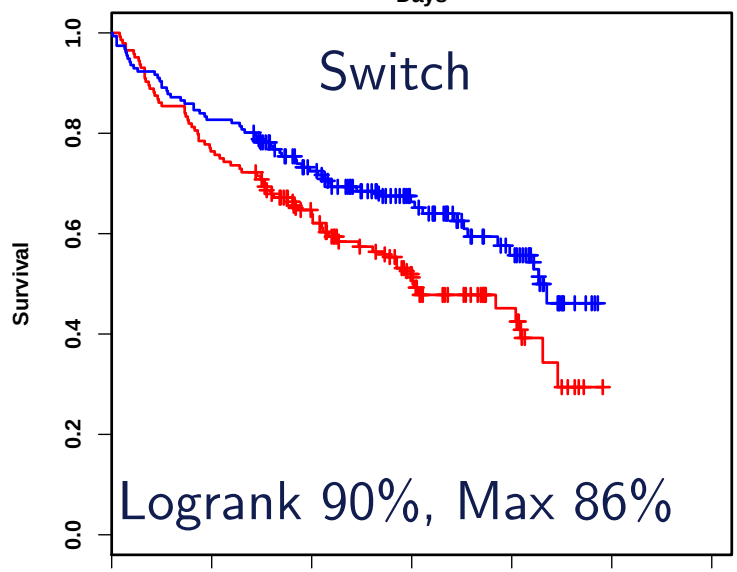


Conditional Power assuming different scenarious









Conditional Power (%) under true scenario is delayed

Test weight (ρ, γ)	Delayed onset	Progression effect	Biomarker Subgroups	Treatment switching
Equal $(0, 0)$	92	97	96	90
Early $(1, 0)$	83	87	87	69
Intermediate $(1, 1)$	90	94	93	80
Late $(0, 1)$	91	97	97	91
Maximum all	89	95	94	86
Maximum $(0, 0)$ $(0, 1)$	89	95	94	85

- Take Conditional Power under scenario which seems most reasonable based on observed interim data
- Alternatively: Give prior on scenarios, calculate posterior and average CP accordingly
- Potential Adaptations: Stopping for futility.
- If study is not stopped for futility, think of further adaptations: change recruitment speed (e.g., decrease to have more late events), sample size and number of events for final analysis
- Interim efficacy testing with group sequential designs Ghosh et al. 2018, Magirr and Jimenez 22

Discussion

- The effect of different sources of non-proportional hazards in clinical trials can be modelled in a piecewise constant hazard framework
- Allows for complex scenarios
- Power simulations to aid planning of trials under assumed non-proportional hazards
- Maximum test as robust alternative to single weighted logrank tests
- Conditional power analysis allows for futility stopping decision
- Note: recruitment and censoring rates, too, determine if more early or late events are observed and hence affect power under non-proportional hazards.
- Implementation of all methods in our R package NPH
- How to quantify effect sizes under NPH? Use simultaneous test of multiple parameters

Ristl et al. 22

Thank you for your attention!

Literature

- Valsamo Anagnostou, Mark Yarchoan, Aaron R Hansen, Hao Wang, Franco Verde, Elad Sharon, Deborah Collyar, Laura QM Chow, and Patrick M Forde. Immuno-oncology trial endpoints: capturing clinically meaningful activity, 2017.
- David Lok Hang Chan, Eva Segelov, Rachel SH Wong, Annabel Smith, Rebecca A Herbertson, Bob T Li, Niall Tebbutt, Timothy Price, and Nick Pavlakis. Epidermal growth factor receptor (egfr) inhibitors for metastatic colorectal cancer. Cochrane Database of Systematic Reviews, (6), 2017.
- Thomas R Fleming and David P Harrington. Counting processes and survival analysis, volume 169. John Wiley & Sons, 2011.
- Ghosh, Ristl, Koenig, Posch, Jennison, Goette, Schueler amd Mehta (2022). Robust group sequential designs for trials with survival endpoints and delayed response. Biometrical Journal, 64(2), 343-360.
- Takahiro Hasegawa. Sample size determination for the weighted log-rank test with the Feming-Harrington class of weights in cancer vaccine studies. Pharmaceutical statistics, 13 (2), 2014.
- Edward L Korn and Boris Freidlin. Interim futility monitoring assessing immune therapies with a potentially delayed treatment effect. Journal of Clinical Oncology, 36(23):2444-2449, 2018.
- Nicholas R Latimer, KR Abrams, PC Lambert, MJ Crowther, AJ Wailoo, JP Morden, RL Akehurst, and MJ Campbell. Adjusting for treatment switching in randomised controlled trials - a simulation study and a simpli
ed two-stage method. Statistical methods in medical research, 26(2):724-751, 2017.
- Magirr and Burman (2019). Modestly weighted logrank tests. Statistics in medicine, 38(20), 3782-3790.
- Richard Peto. Rank tests of maximal power against Lehmann-type alternatives. Biometrika, 59(2):472-475, 1972.
- Ristl, Ballarini, Goette, Schueler, Posch, and Koenig, F. (2021). Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze RCTs in oncology. Pharmaceutical Statistics, 20(1), 129-145.
- David Schoenfeld. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika, 68(1):316-319, 1981.
- Robert E Tarone. On the distribution of the maximum of the logrank statistic and the modi
ed Wilcoxon statistic. Biometrics, pages 79-85, 1981.

Backup: Piecewise constant hazard function

For each subpopulation and treatment group, define survival function via a piecewise constant hazard function.

Define k time intervals $[t_{i-1}, t_i)$ with $0 = t_0 < t_1 \ldots < t_k = \infty$ and constant hazard λ_i.

The hazard function is $\lambda(t) = \sum_{i=1}^{k} \lambda_i \mathbb{1}_{t \in [t_i - 1, t_i)}$

And the survival function is

$$S(t) = \exp\left\{ \int_0^t \lambda(s) ds \right\}$$

Backup: Model different hazard after disease progression

X ... time to disease progression
$\lambda_D(t)$ ... hazard function before disease progression
$\lambda_{PD}(t)$ ... hazard function after disease progression
Conditional on $X = u$, the hazard function and survival function are

$$\lambda(t|X = u) = \lambda_P(t)\mathbb{1}_{t \leq u} + \lambda_{PD}(t)\mathbb{1}_{t > u}$$

$$S(t|X = u) = \exp\left\{-\int_0^t \lambda(s|X = u)ds\right\}$$

Distribution of X is modeled via piecewise constant hazards, independent from hazard functions $\lambda_P(t)$ and $\lambda_{PD}(t)$.
The survival function is obtained by integrating over the possible progression time points

$$S(t) = \int_0^t S(t|X = u)p(X = u)du$$

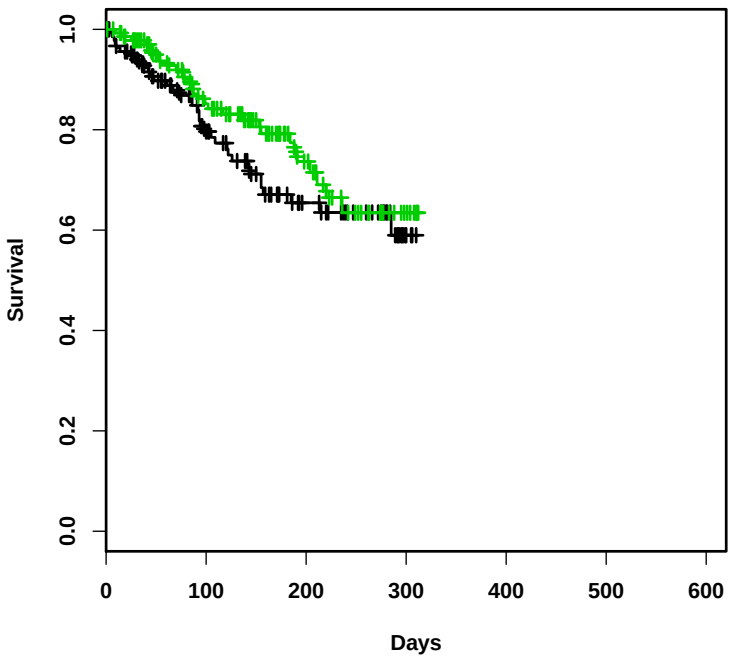
Note: Correlation between progression free survival (PFS) and overall survival (OS) results from $PFS = \min(OS, D)$ and from different hazards

Backup: Mixture of subpopulations

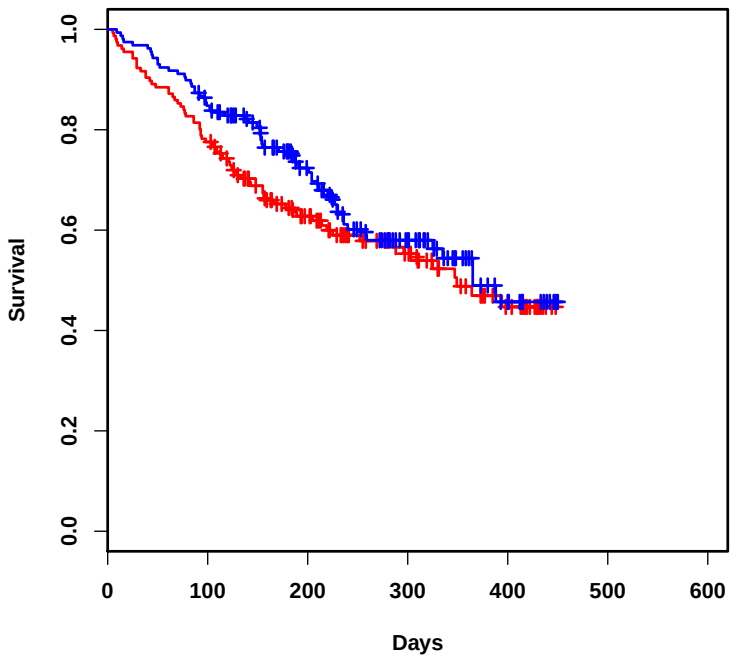
- In general, assume m subpopulations with proportions $p_i, i = 1, \ldots, m$ of the full population and survival functions S_i.
- Then the overall population survival function is $S(t) = \sum_{i=1}^{m} p_i S_i(t)$.

Conditional data simulation Case Study B

Observed Interim

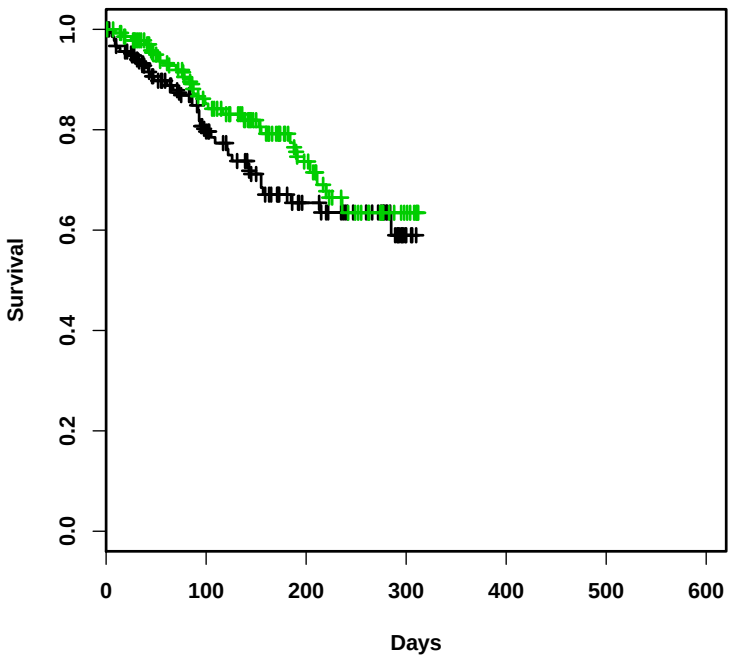


Simulated Delayed

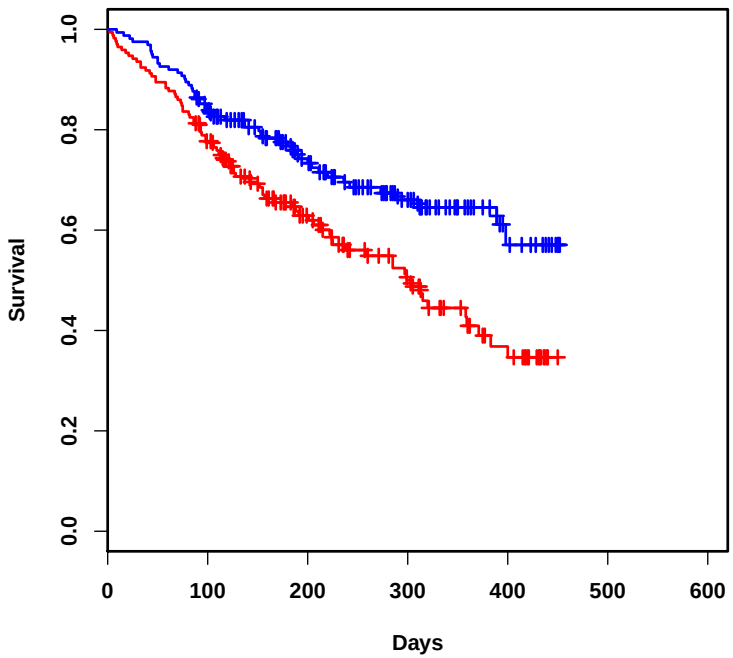


Conditional data simulation Case Study B

Observed Interim

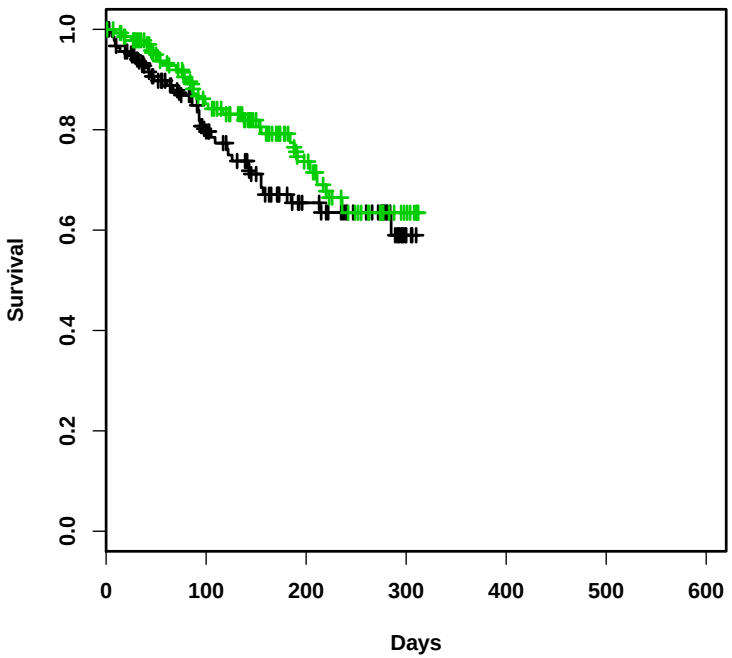


Simulated Delayed

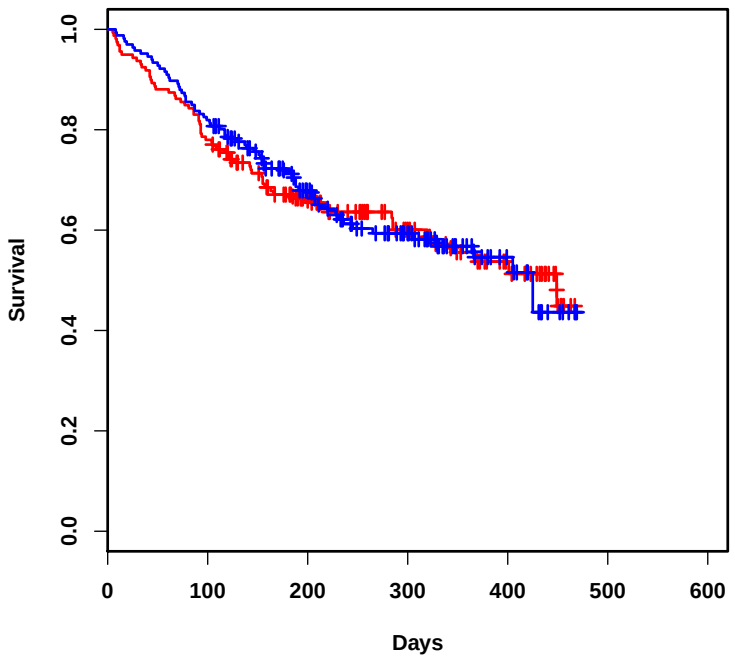


Conditional data simulation Case Study B

Observed Interim

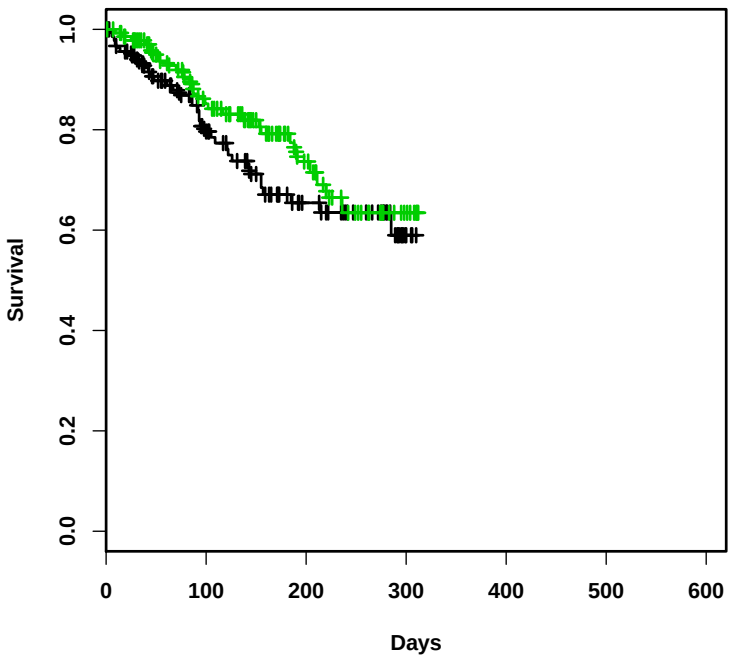


Simulated Delayed

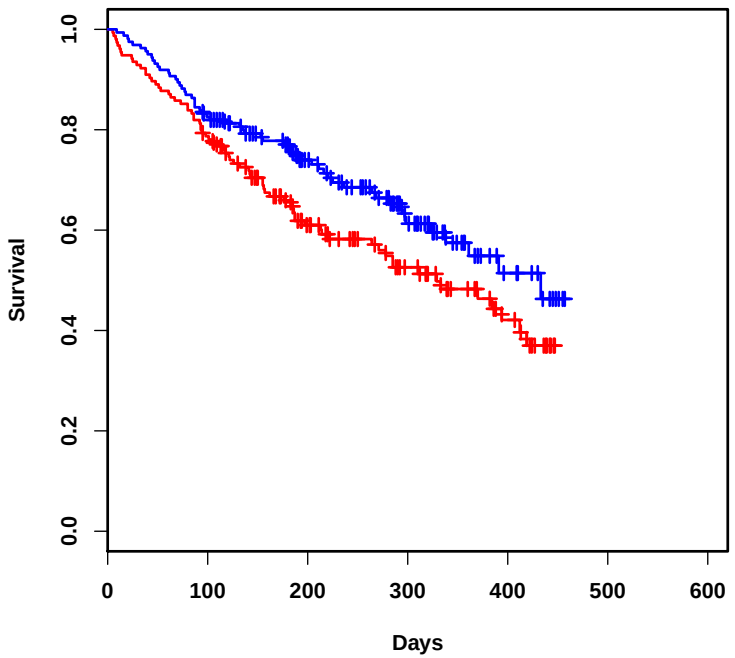


Conditional data simulation Case Study B

Observed Interim

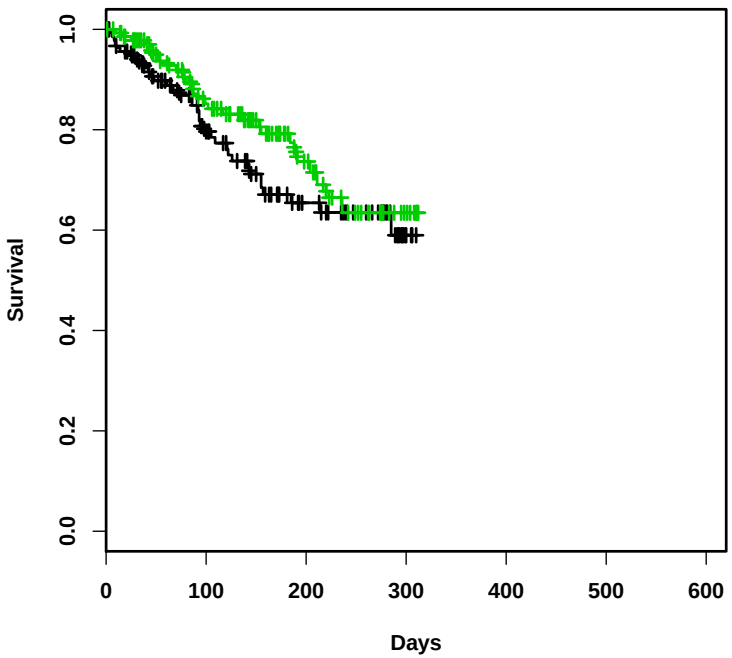


Simulated Delayed

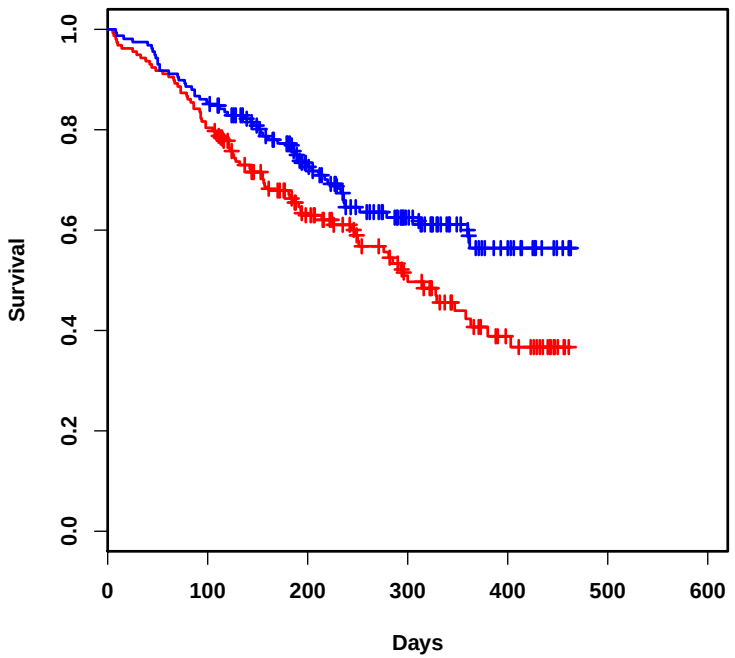


Conditional data simulation Case Study B

Observed Interim

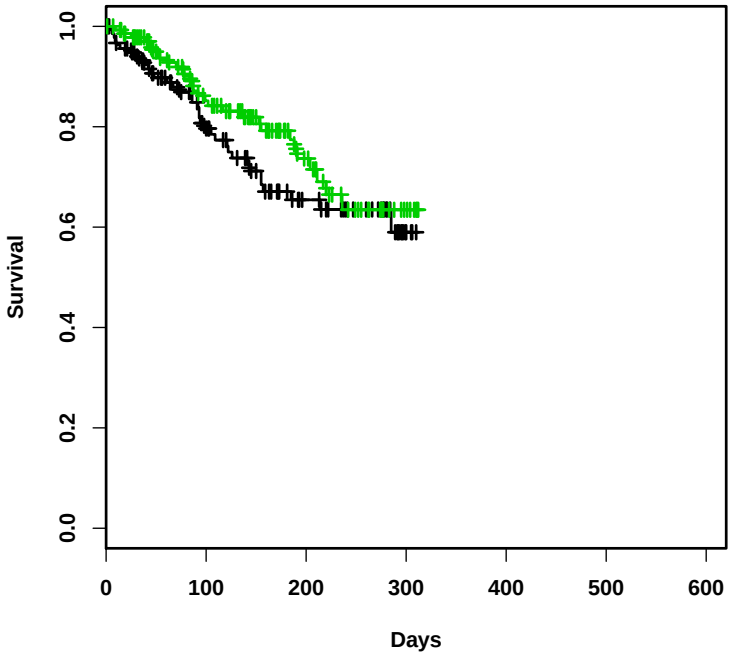


Simulated Delayed

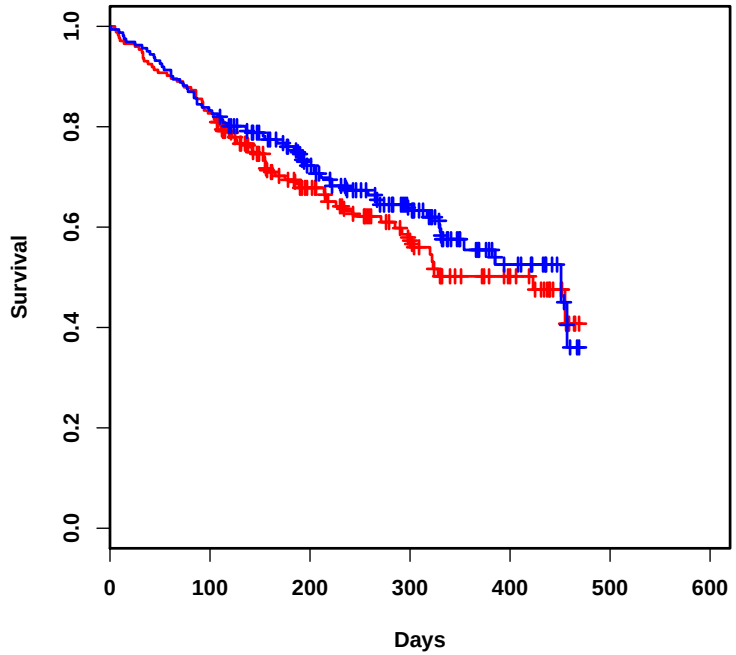


Conditional data simulation Case Study B

Observed Interim

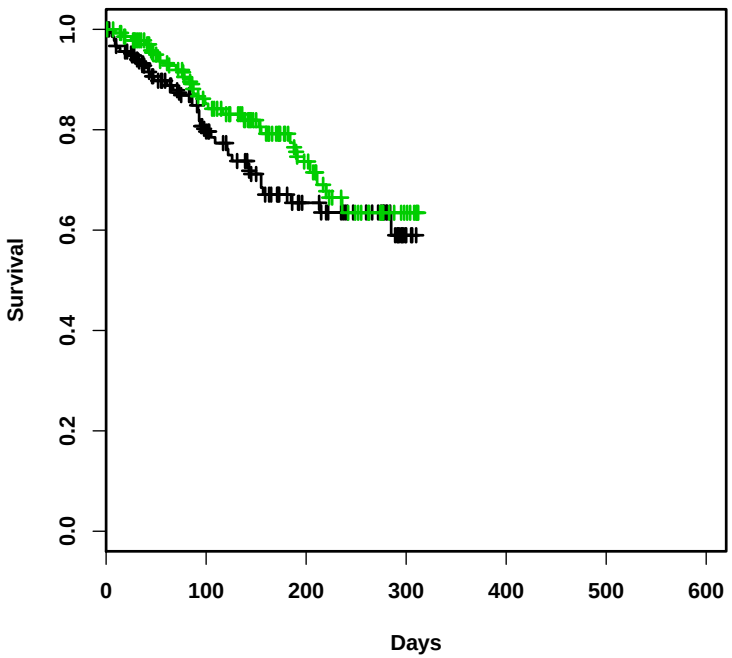


Simulated Delayed

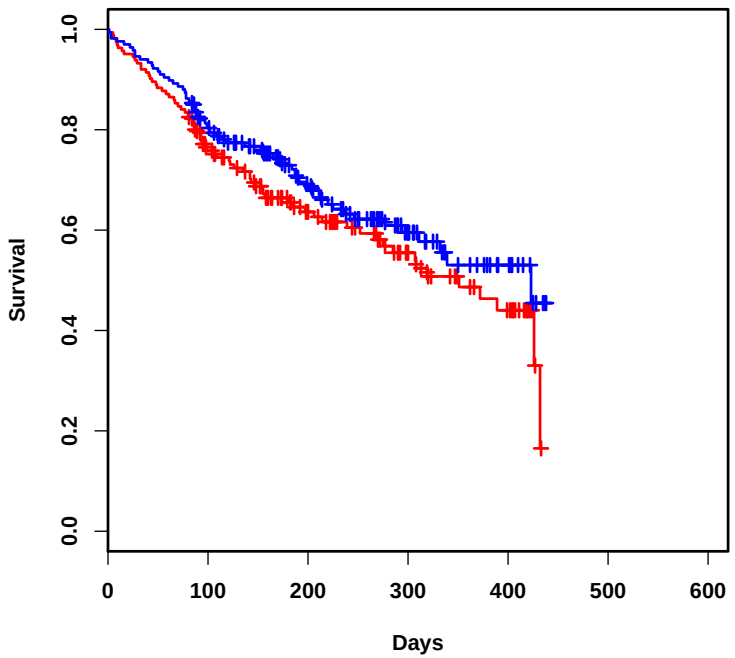


Conditional data simulation Case Study B

Observed Interim

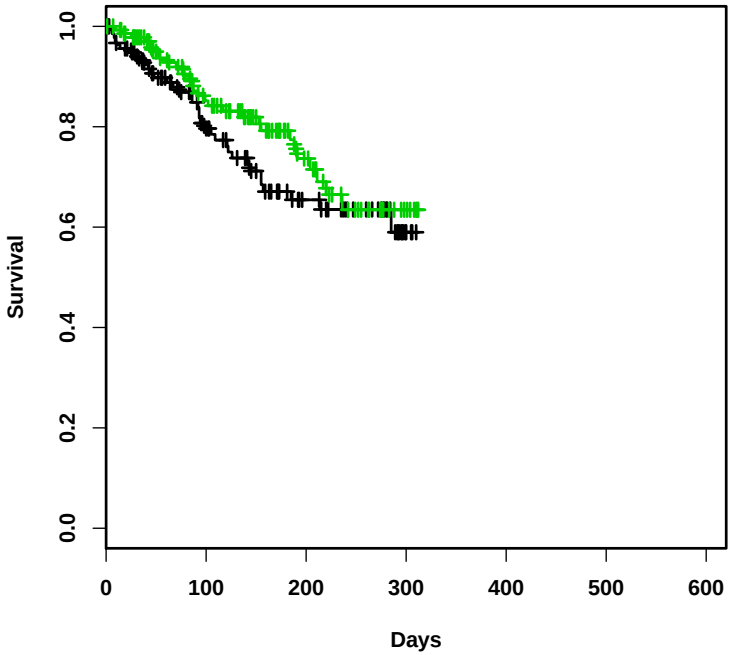


Simulated Delayed

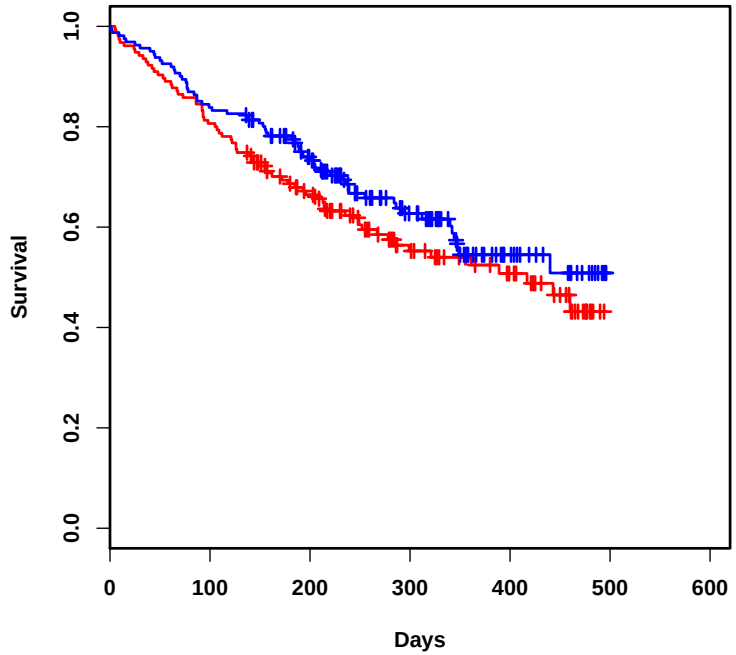


Conditional data simulation Case Study B

Observed Interim

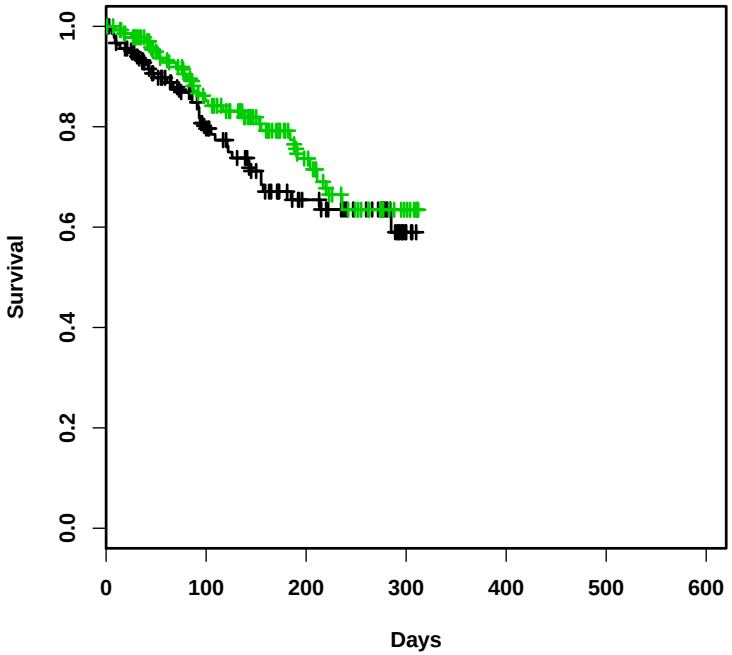


Simulated Delayed

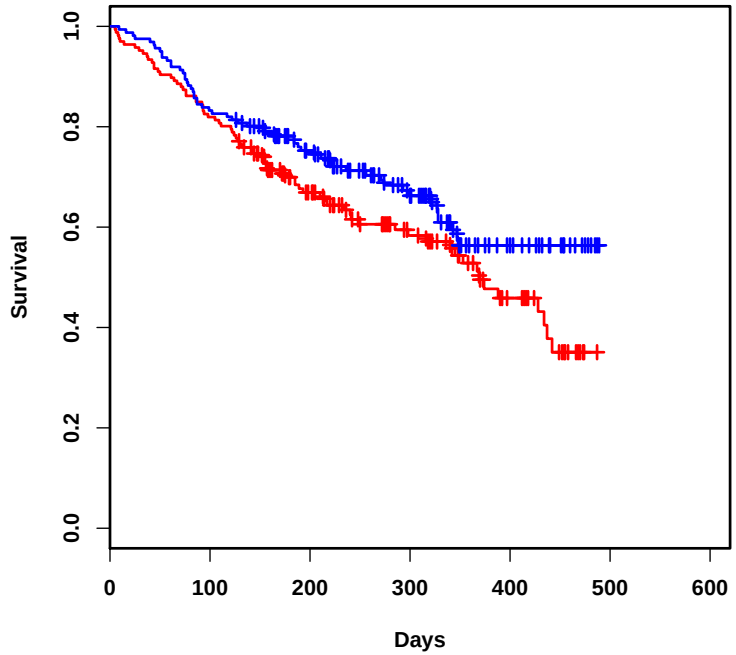


Conditional data simulation Case Study B

Observed Interim

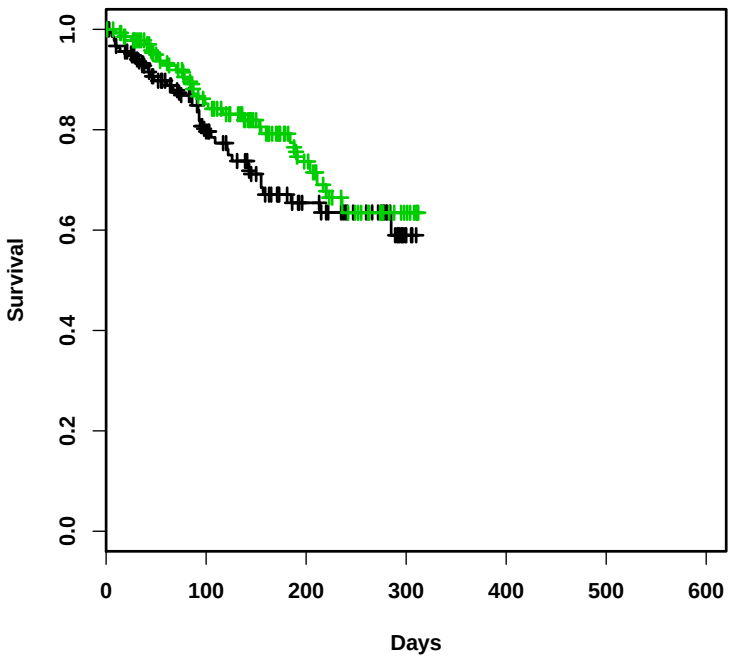


Simulated Delayed

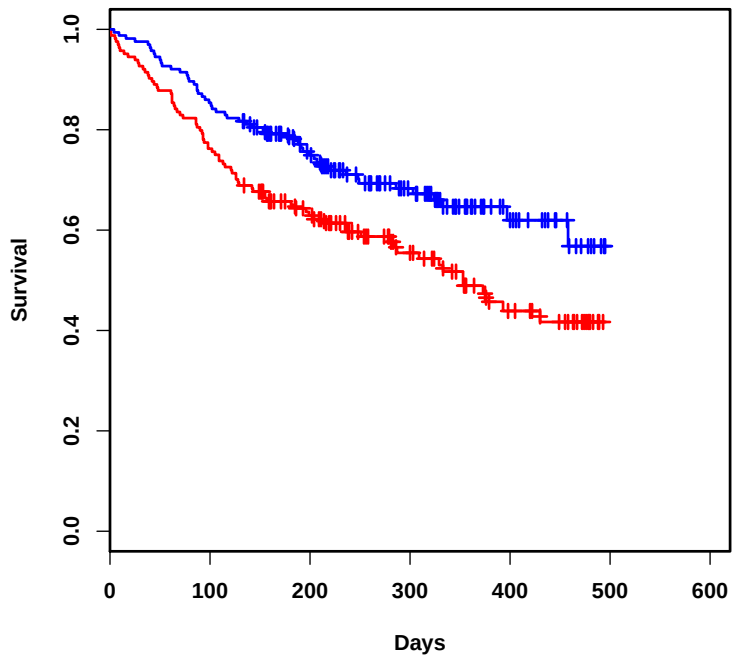


Conditional data simulation Case Study B

Observed Interim



Simulated Delayed



Conditional power Logrank test 44%, Maximum test 36%

Conditional Power (%) for Case Study B

Test weight (ρ, γ)	Delayed onset	Progression effect	Biomarker Subgroups	Treatment switching
Equal $(0, 0)$	44	62	62	44
Early $(1, 0)$	30	33	38	19
Intermediate $(1, 1)$	31	37	40	21
Late $(0, 1)$	43	66	64	49
Maximum all	36	54	53	36
Maximum $(0, 0)$ $(0, 1)$	37	53	53	34

- Conditional Power less in Case Study B
- For Max test lowest power for delayed onset and treatment switch scenarios
- If study is not stopped for futility, think of further adaptations: change recruitment speed (e.g., decrease to have more late events), sample size and number of events for final analysis

Adaptive Tests for Survival Data (I)

Patients recruited in the first stage maybe still under risk in the second stage.

- Early Stopping for Efficacy: Group sequential designs for maximum tests

  Mehta & Ghosh 2018, Ghosh, 2018

- Adaptions: The combination test and the conditional error approach can be extended to survival data and the (weighted) log-rank test (independent increments property).

  Wassmer 2006, Schaefer & Mueller 2001

- Stagewise p-values are calculated from the events occuring in each stage.
- Caveat: This may lead to biased tests if adaptations are based on covariate information or secondary endpoints of first stage patients censored at the time of the interim analysis. E.g., adaptations based on PFS when the primary endpoint is OS.

  Bauer & Posch, 2001

Adaptive Survival Trials (II)

- Test procedures where the follow-up time from first stage patients is fixed control the type I error rate, but may not include all events in the test statistics if the trial is extended.

  Jenkins et al. '11, Irle & Schäfer, '12, Joergens et al. '19

- Conservative tests based on all observed data are typically strictly conservative.

  Magirr et al. 2016