Confirmative assessment of differences in the survival function based on multiple characteristics

Martin Posch

Medical University of Vienna, Vienna, Austria

joint work with Heiko Götte, Armin Schüler (Merck), Franz König, Robin Ristl (Medical University of Vienna)

PSI One-Day Meeting
Non-proportional hazards and applications in immuno-oncology
April 29, 2021

Which hypothesis tests to use under NPH?

- Logrank/Cox test: loss in power, hazard ratio is not well defined
- Weighted logrank test: gives larger weights to time-points where hazards are expected to strongly differ between groups.
  Fleming and Harrington (2011), Magirr and Burman (2019). Jiménez, et al. (2020).
- Tailored tests based on Brownian motion approximations
  Chauvel and O'Quigley (2014), Flandre and O'Quigley (2019)
- Tests based on differences of survival functions as the two-Sample Tests of Cramér–von–Mises- and Kolmogorov–Smirnov-Type Schumacher (1984)
- ...

Combining several tests to guarantee power in a range of scenarios

If the onset and duration of the treatment effect is unknown several hypothesis tests can be combined:

- Max-Combo test
  Maximum statistics of several weighted log-rank tests

  Tarone (1981), Lee (2007), Karrison (2016), Ristl et al. (2020)
- Combination of distance-from-origin test and area-under-the-curve test

  Chauvel and O'Quigley (2014)
- Combination of log-rank/Cox test and a permutation test for the restricted mean survival time

  Royston and Parmar (2016), Royston (2017), Royston et al. (2019)
- ...

Criticism of weighted tests

- General weighted tests, test the null hypothesis

$$H_0 : S_1(t) = S_0(t) \text{ for all } t.$$

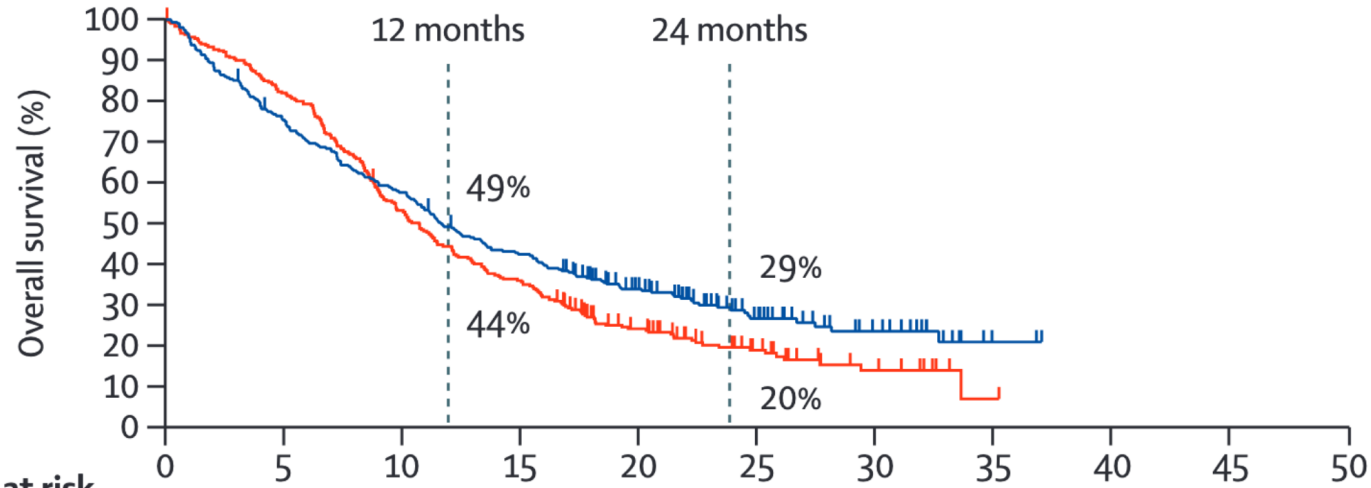
- The modestly weighted test, tests the null hypothesis

$$H_0 : S_1(t) \leq S_0(t) \text{ for all } t.$$

- In both cases, rejection of the null hypothesis only implies that the curves are not equal but does not exclude crossing survival curves.
- To maximize power, more weight is given to regions of the survival curves, where an effect in the desired direction is anticipated.
- Rejection does not imply that the novel treatment has a superior survival.

Magirr and Burman (2019), Bartlett et al. (2020)

Pembrolizumab vs Cetuximab in squamous cell carcinoma of the head and neck (KEYNOTE-048)



Number at risk (number censored)											
Pembrolizumab alone	301 (0)	225 (2)	172 (2)	125 (4)	81 (24)	37 (55)	18 (71)	2 (86)	0 (88)	0 (88)	0 (88)
Cetuximab with chemotherapy	300 (0)	245 (1)	158 (2)	107 (2)	57 (19)	26 (40)	10 (51)	1 (59)	0 (60)	0 (60)	0 (60)

Pembrolizumab alone (blue) vs Cetuximab w chemotherapy (red), total population

Burtness et al. (2020)

Treatment choice in case of crossing survival curves

- For crossing survival curves, which survival distribution is more desirable?
- What is the weight patients and physicians give to different regions of the survival curve?
- Which characteristics of the survival curves are relevant?
- Preference studies elicited the utilities of different survival time distributions for patients and physicians

Shafrin et al. (2017), Hauber et al. (2020)

Fixed or variable but potentially durable survival?

Instructions:

Think about a patient who develops advanced melanoma, and whose cancer progresses after receiving one kind of therapy for their advanced disease.

Suppose there were two additional therapies they could try next. Let's call these hypothetical therapies "Therapy A" and "Therapy B."

Therapy A and Therapy B are administered in the same way, and have the same minimal side effects.

With BOTH therapies, an AVERAGE patient can expect to live 4 YEARS.

However, the therapies are not completely the same in how long patients live.

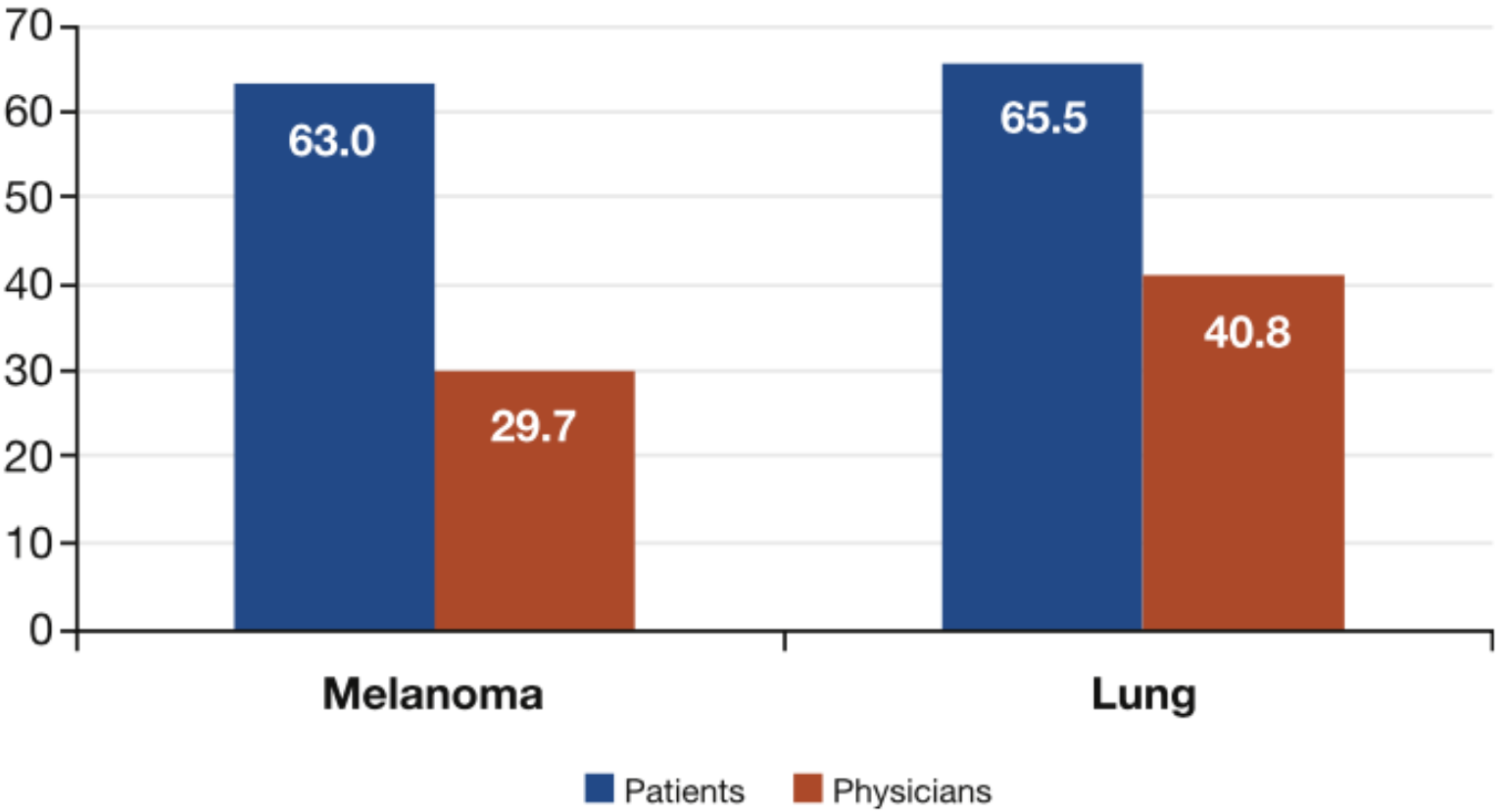
Therapy A	**Therapy B**
In particular, with THERAPY A, nearly ALL patients live EXACTLY 4 YEARS, no more and no less.	However, with THERAPY B, • HALF (50%) of patients live 1 YEAR OR LESS • THREE OUT OF TEN (30%) patients live MORE THAN 1 YEAR but LESS THAN 7 YEARS • TWO OUT OF TEN (20%) patients live 7 YEARS OR MORE

If the therapies required the same out-of-pocket cost to the patient, which would you personally prefer?

○ **Therapy A**
○ **Therapy B**

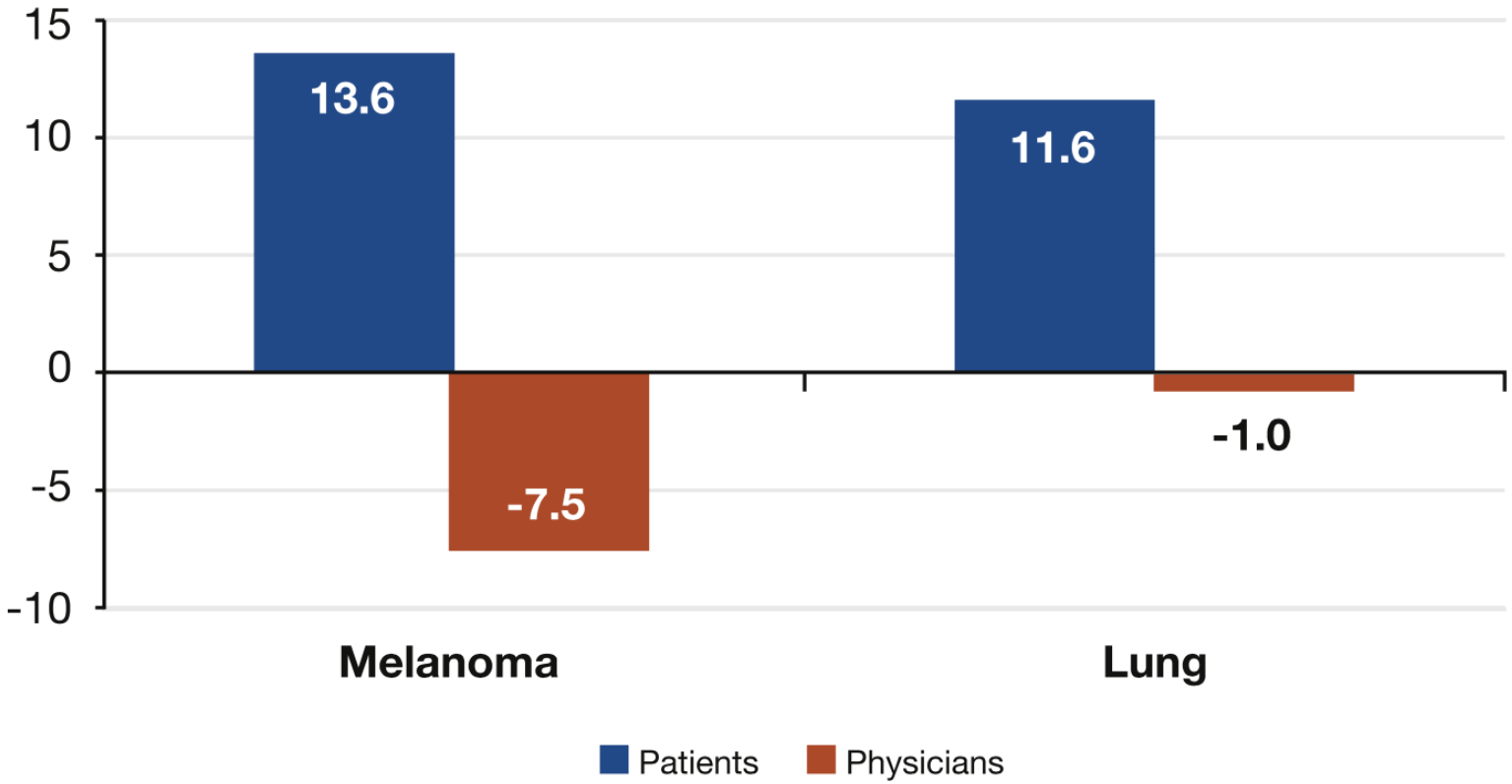
Shafrin et al. (2017)

Proportion choosing Therapy B (with 20% long term survival)

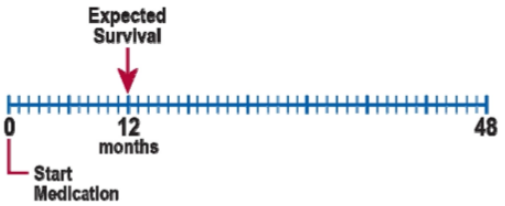
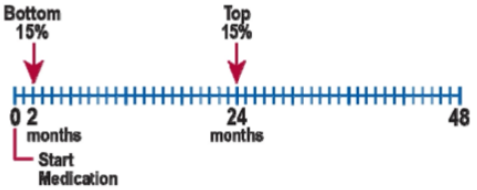
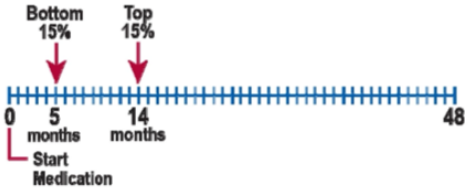


Advanced stage melanoma ($n = 81$) and lung cancer ($n = 84$) patients and oncologists ($n = 91/96$). Shafrin et al. (2017)

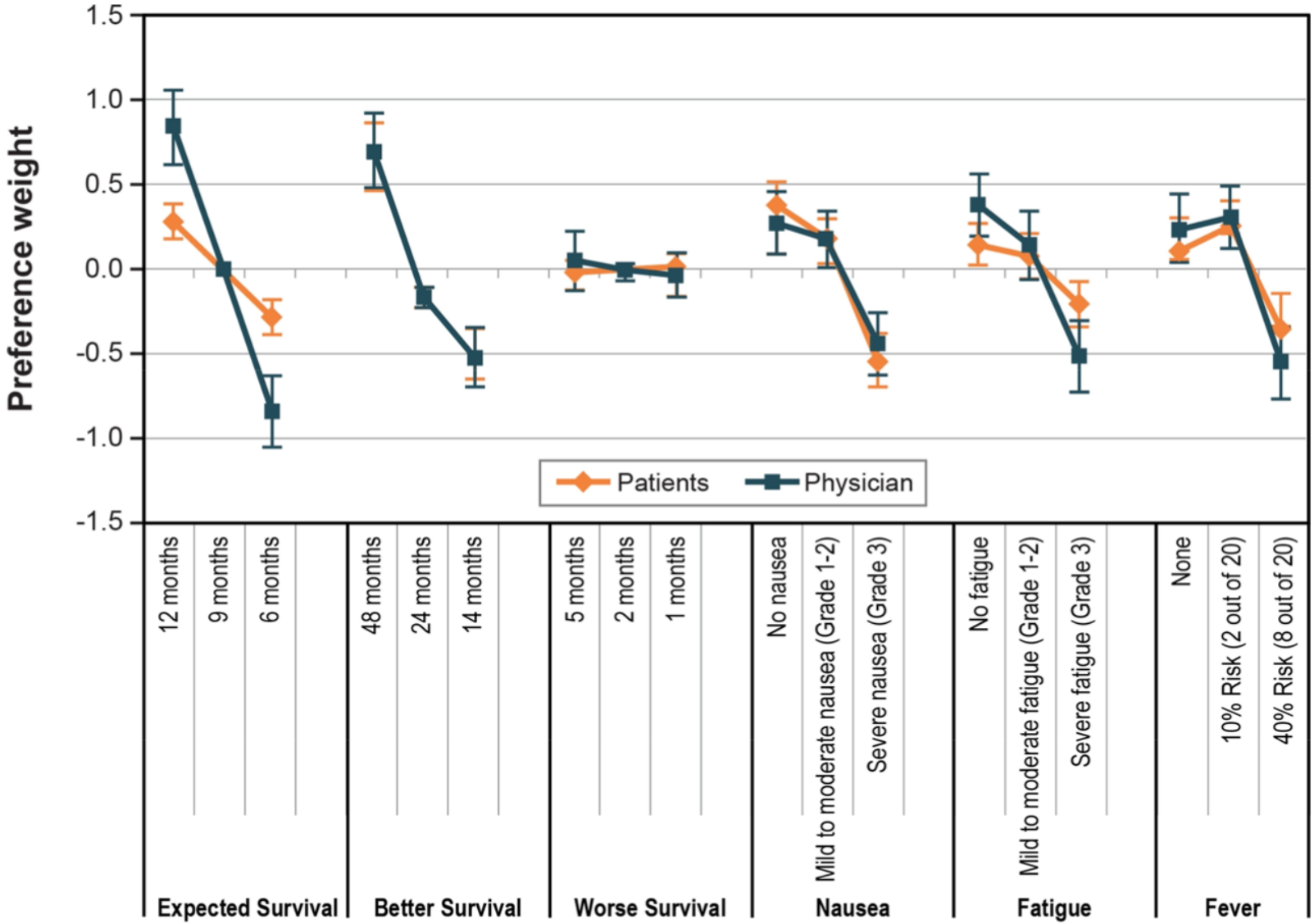
Months of additional survival needed from therapy with fixed survival to be indifferent to Therapy B



Shafrin et al. (2017)

Treatment Feature	Treatment A	Treatment B
Expected survival with treatment		
Survival for bottom 15% of patients Survival for top 15% of patients		
Fatigue	No fatigue	Mild to moderate fatigue (Grades 1 and 2)
Nausea	Mild to moderate nausea (Grades 1 and 2)	No nausea
Risk of febrile neutropenia	 10% (2 out of 20)	 40% (8 out of 20)
Which would you choose	☐	☐

Hauber et al. (2020)



200 advanced non-small cell lung cancer patients and 100 oncologists.

Hauber et al. (2020)

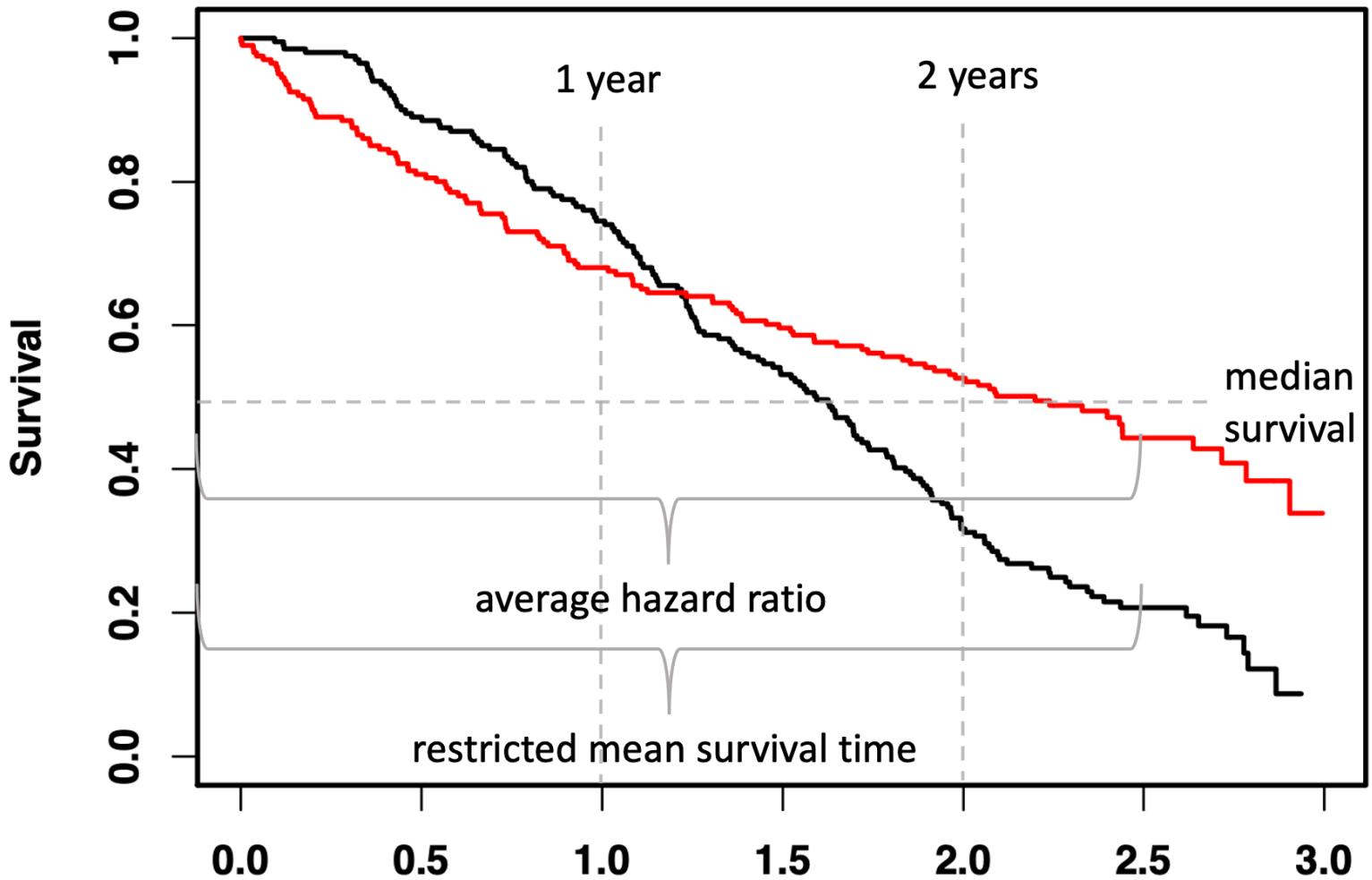
What are the relevant measures of treatment effects?

- Utility is based on trade-offs between short-term, expected and long-term survival
- Studies show that patients appear to be more risk seeking than physicians
- However, there is a large variability in the preferences within patients (and within physicians).
- Also for regulators and payers several aspects of the survival curves are relevant for decision making.

How to quantify effect sizes under NPH?

- **Landmark analysis** (e.g., Differences in 1-year, 2-year survival).
- **Differences in Restricted Mean Survival Time**: measure of average survival from time 0 to a specified time point. Irwin (1949)
- **Average Hazard Ratio**: from time 0 to a specified time point; interpretation akin to Mann-Whitney statistic.
  Kalbfleisch & Prentice (1981), Brückner & Brannath (2017), Rauch et al.(2018)
- **Difference in Median Survival**
- Parametric models (e.g. three-component cure rate model)
  Kim and Gray (2012)
- Simultaneous confidence bands for difference of survival functions
  Parzen (1997)

Proposal: Simultaneous test of multiple parameters



Proposal: Simultaneous test of multiple parameters

- Quantify the effect size by
  - 1-year survival difference, θ_1
  - 2-year survival difference, θ_2
  - Median survival difference, θ_3
  - Average hazard ratio, θ_4
  - Difference in restricted mean survival time θ_5
  - ...
- Tests of the null hypotheses

$$H_j : \theta_j = \theta_{0,j}, \quad j = 1, \ldots, m$$

  controlling the family wise error rate (FWER) at a level α.
- Simultaneous confidence intervals for all considered parameters with overall coverage probability $1 - \alpha$.

Multiple testing & simultaneous confidence intervals

- Bonferroni adjustment is very conservative if the parameter estimates are highly correlated.
- The Bonferroni test can be improved by accounting for the correlation
- To derive the critical values we use a multivariate normal approximation of the distribution of the estimates

$$(\hat{\theta}_1, \ldots, \hat{\theta}_m)$$

based on the asymptotic covariance matrix.

Derivation of the asymptotic covariance matrix

- The difference between observed $N_i(t)$ and expected number of events is a martingale

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) d\Lambda_i(s),$$

where $Y_i(t)$ is the number under risk and $\Lambda_i(t)$ the cumulative hazard function

Aalen (2010), Fleming and Harrington (1991)

- All considered estimators can be written as stochastic integrals with dM_i

$$\hat{\theta}_{k,i} - \theta_{k,i} = a_{k,i} \int_0^{t_k} H_{k,i}(s) \frac{1}{Y_i(s)} dM_i(s) + o_p(1/n_i)$$

  - $a_{k,i}$ is a constant or a parameter with a consistent estimate $\hat{a}_{k,i}$
  - $H_{k,i}$ is a predictable process with respect to M_i

Application of multivariate central limit theorem

- True parameters are of the form $\theta_k = \theta_{k,1} - \theta_{k,0}$
- By the martingale representation, the estimators in each group asymptotically follow a multivariate normal distribution.
  Fleming and Harrington (1991)
- A consistent estimate for their covariance matrix $\boldsymbol{\Sigma}_i$ has elements

$$\hat{\mathrm{cov}}(\hat{\theta}_{k,i}, \hat{\theta}_{k',i}) = \hat{a}_{k,i}\hat{a}_{k',i} \sum_{s \in D_i, s \leq t_k \wedge t_{k'}} H_{k,i}(s) H_{k',i}(s) \frac{1}{Y_i^2(s)} dN_i(s),$$

  where D_i is the set of observed event times in group i.
- Hence, $(\hat{\theta}_1, \ldots, \hat{\theta}_m)$ can be approximated by a multivariate normal distribution with mean $(\theta_1, \ldots, \theta_m)$ and covariance matrix $\hat{\boldsymbol{\Sigma}} = \hat{\boldsymbol{\Sigma}}_0 + \hat{\boldsymbol{\Sigma}}_1$.

Selected elements of the covariance matrix estimator

	$\log \hat{S}_i(t)$	$\log \hat{q}_{\gamma,i}$	$\log avrHR(L)$
$\log \hat{S}_i(t)$	$\sum_{s \le t} \frac{dN_i(s)}{Y_i^2(s)}$	$\frac{-1}{\hat{q}_{\gamma,i}\hat{\lambda}(\hat{q}_{\gamma,i})} \sum_{s \le t \wedge \hat{q}_{\gamma,i}} \frac{dN_i(s)}{Y_i^2(s)}$	$\frac{-1}{\hat{\pi}_i(L)} \sum_{s \le t \wedge L} (\hat{S}_0\hat{S}_1)(s) \frac{dN_i(s)}{Y_i^2(s)}$
$\log \hat{q}_{\gamma,i}$		$\frac{1}{\hat{q}^2_{\gamma,i}\hat{\lambda}^2(\hat{q}_{\gamma,i})} \sum_{s \le \hat{q}_{\gamma,i}} \frac{dN_i(s)}{Y_i^2(s)}$	$\frac{1}{\hat{q}_{\gamma,i}\hat{\lambda}(\hat{q}_{\gamma,i})\hat{\pi}_i(L)} \sum_{s \le \hat{q}_{\gamma,i} \wedge L} (\hat{S}_0\hat{S}_1)(s) \frac{dN_i(s)}{Y_i^2(s)}$
$\log avrHR(L)$			$\frac{1}{\hat{\pi}_i^2(L)} \sum_{s \le L} (\hat{S}_0\hat{S}_1)^2(s) \frac{dN_i(s)}{Y_i^2(s)}$
$\log \hat{S}(t')$	$\sum_{s \le t \wedge t'} \frac{dN_i(s)}{Y_i^2(s)}$		

All sums are restricted to the set of observed event times $s \in \mathcal{D}_i$.

Perturbation approach to estimate covariance matrix

A "parametric bootstrap" type approach based on the asymptotically normality of the independent increments of the Nelson-Aalen estimator of cumulative hazards.

1. Generate random increments to generate "perturbations" of the observed cumulative hazard estimates in the treatment and the control group and calculate the considered parameter estimates
2. Repeat (1.) a large number of times
3. Calculate their empirical covariance matrix.

See Park and Wei (2003), Zhao et al. (2016) for other applications of this approach.

Testing procedure and confidence intervals

- For a vector $\hat{\boldsymbol{\theta}} \sim N_m(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ of parameter estimates consider the test statistics

$$\boldsymbol{T}_j = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)/\sqrt{\hat{\sigma}_j^2},$$

- Under $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, $\boldsymbol{T}(\boldsymbol{\theta}_0)$ is multivariate normal distributed.
- H_j is rejected if $\boldsymbol{T}_j \geq c$, where c is chosen such that

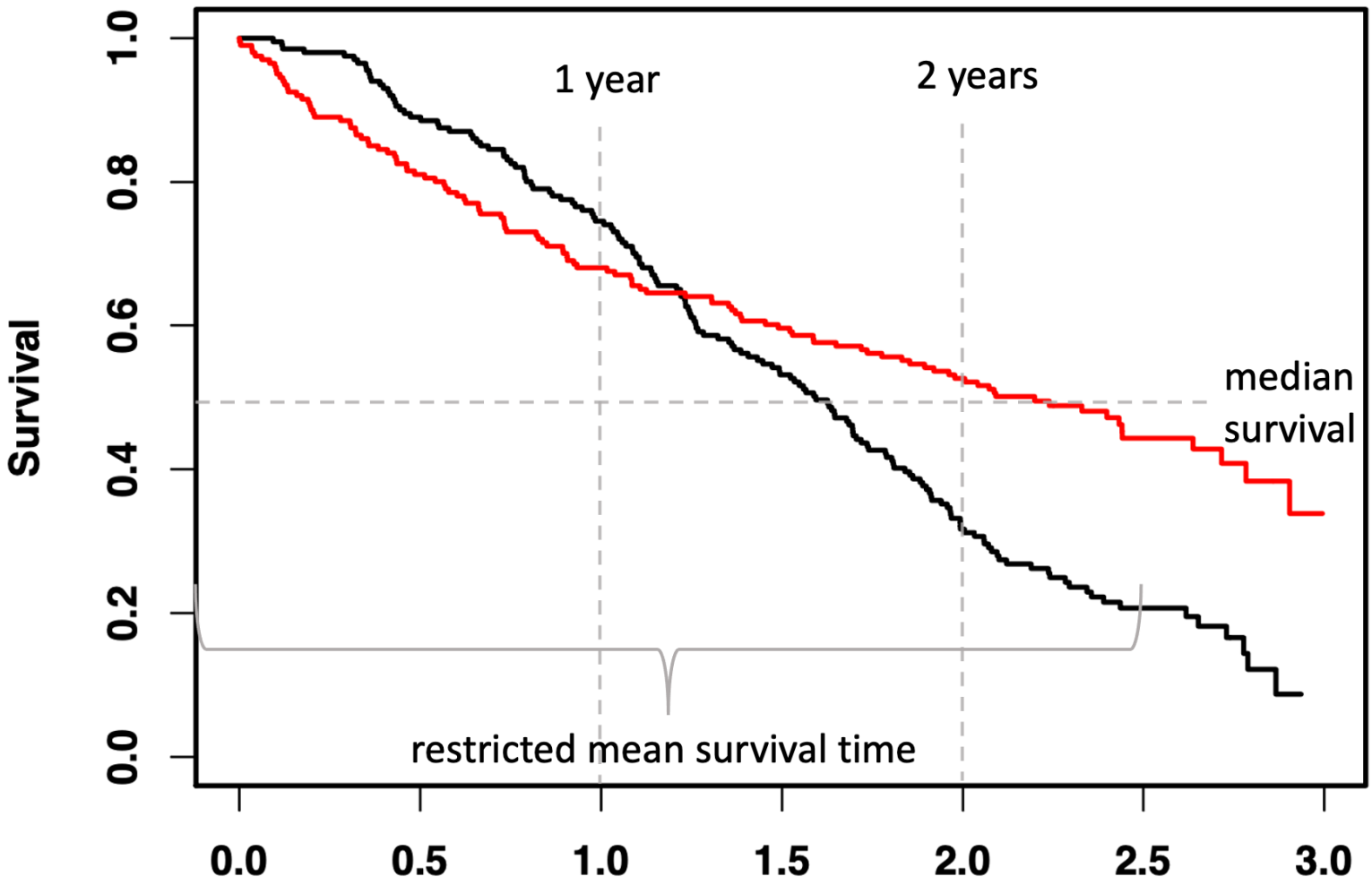
$$P(\max_{i=1,\ldots,m} \boldsymbol{T}_i \geq c) = \alpha/2.$$

- A further improved can be a achieved with the closed testing procedure.

Hothorn, Bretz and Westfall (2008)

- Simultaneous confidence intervals are given by $\hat{\theta}_j \pm \sqrt{\hat{\sigma}_j^2} c$

Numerical example



Numerical example

Parameter	Estimate (SE)		95% CI	p-value
1-year Survival	-0.06 (0.05)	u	(-0.15,0.02)	0.15
		a	(-0.17,0.04)	0.31
2-year Survival	0.21 (0.05)	u	(0.12, 0.30)	< 0.001
		a	(0.10,0.32)	< 0.001
Median Survival	0.61 (0.23)	u	(0.16,1.06)	0.01
		a	(0.07,1.14)	0.02
RMST (2.5)	0.11 (0.08)	u	(-0.05,0.28)	0.18
		a	(-0.08,0.31)	0.36

Simulation study

Considered parameters:

- Log ratio of survival probabilities for two or three time points
- Log ratio for one selected quantile
- Average hazard ratio over the maximal observation time span

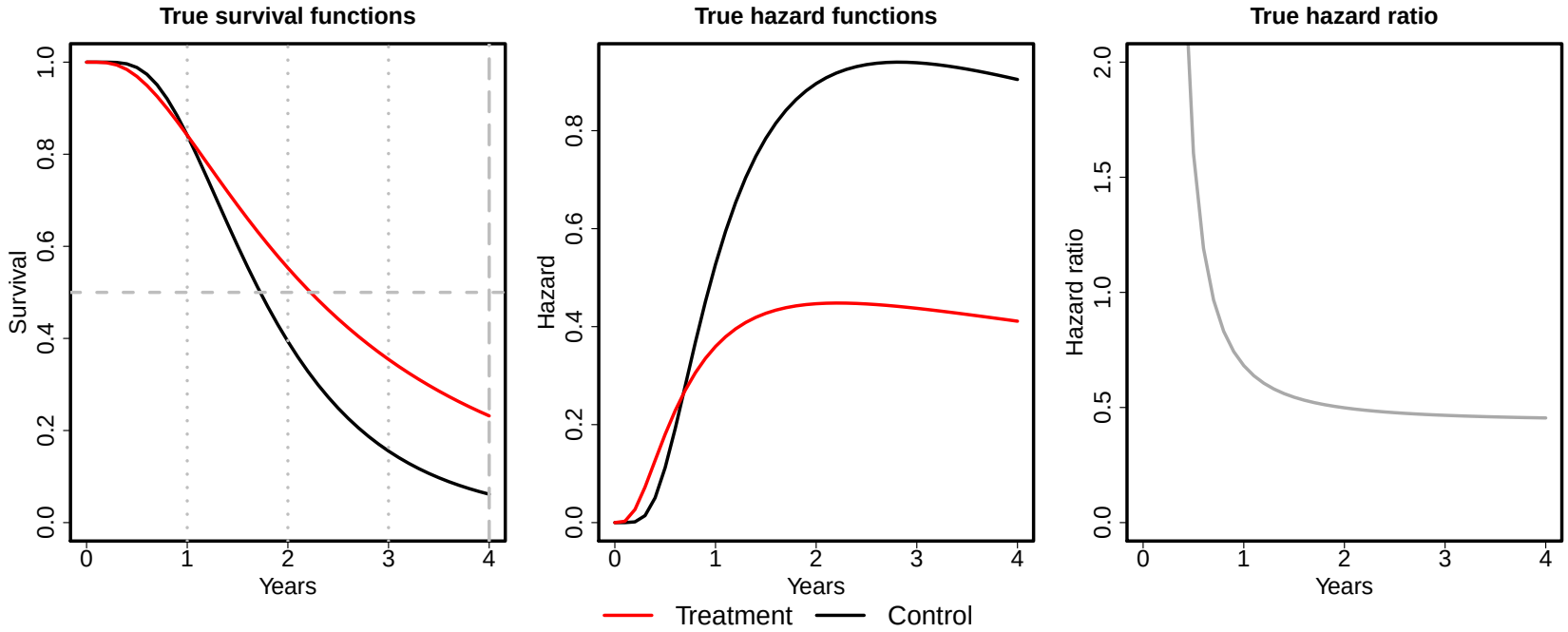
Simulation settings

- 50,000 simulation runs
- Equal sample size per group of 75, 100, 150, 250

Inference procedures in the simulation

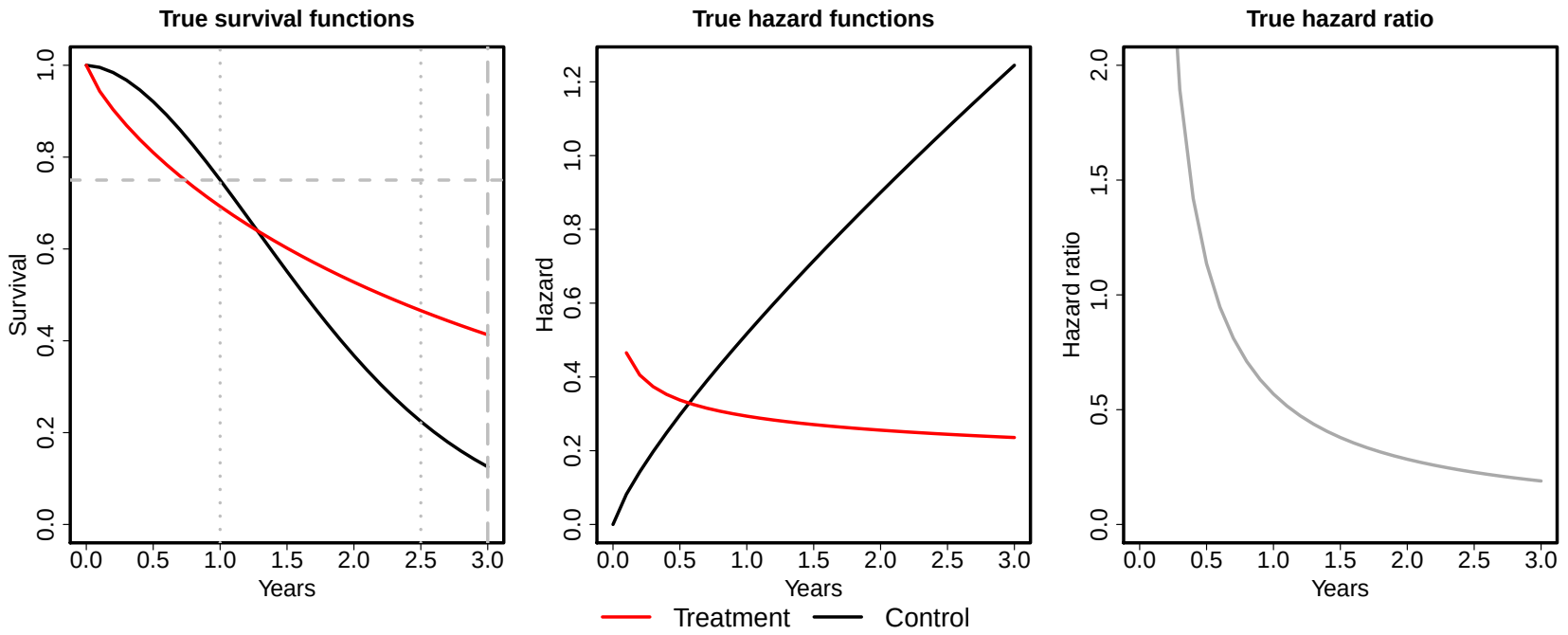
- Test for individual H_j at multiple level 2.5%
- Simultaneous 95% confidence intervals using multivariate normal approximation
- Asymptotic covariance matrix estimate
- Perturbation covariance matrix
- Bonferroni test and Bonferroni-adjusted 95% confidence intervals

Scenario 1: Delayed onset of treatment effect



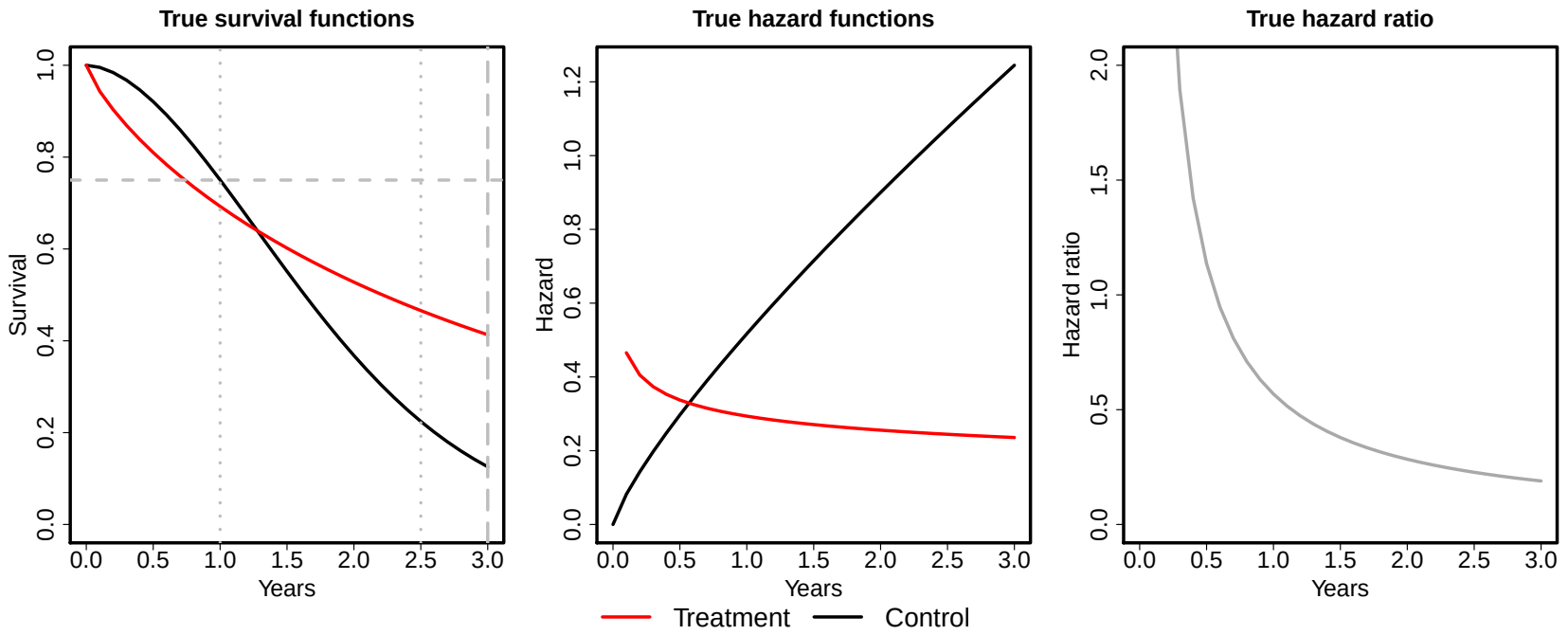
- Modelled via lognormal distributions
- Two years recruitment, four years total study duration
- Further random censoring times distributed as lognormal(4,9).
- 56% vs. 72% of patients with event

Scenario 2: Crossing survival, fast recruitment



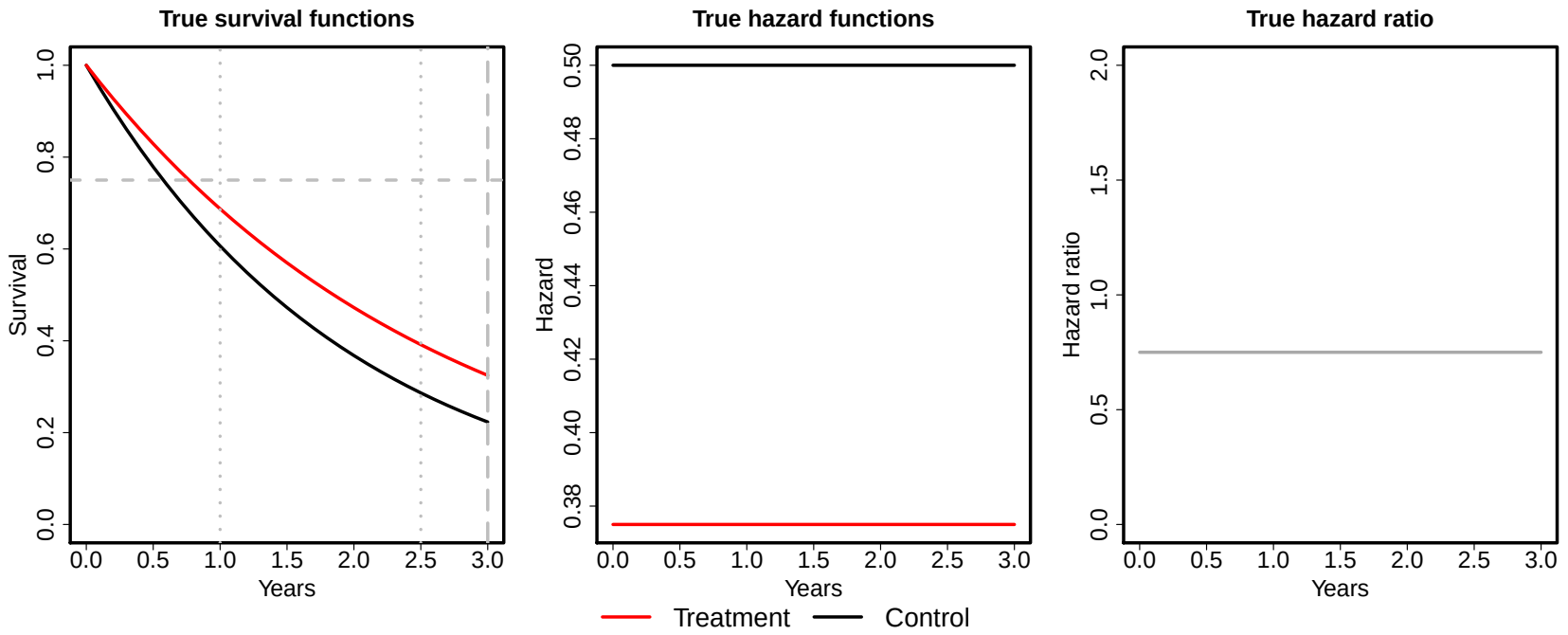
- Modelled via Weibull distributions
- One year recruitment, three years total study duration
- 53% vs. 77% of patients with event

Scenario 3: Crossing survival, slow recruitment



- Modelled via Weibull distributions
- Two years recruitment, three years total study duration
- 47% vs. 59% of patients with event

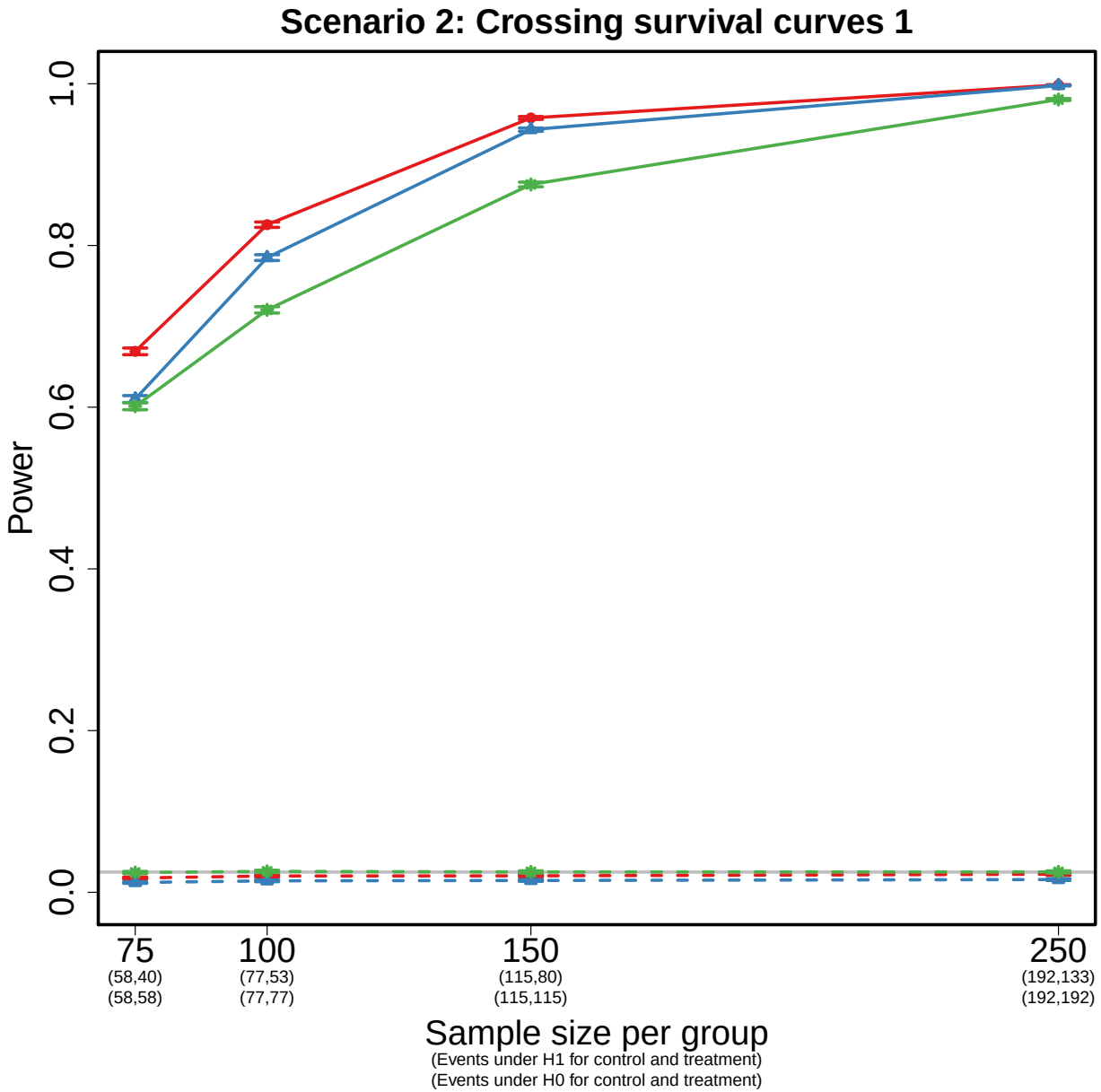
Scenario 4: Proportional hazards



- Modelled via Exponential distributions
- One year recruitment, three years total study duration
- 61% vs. 70% of patients with event

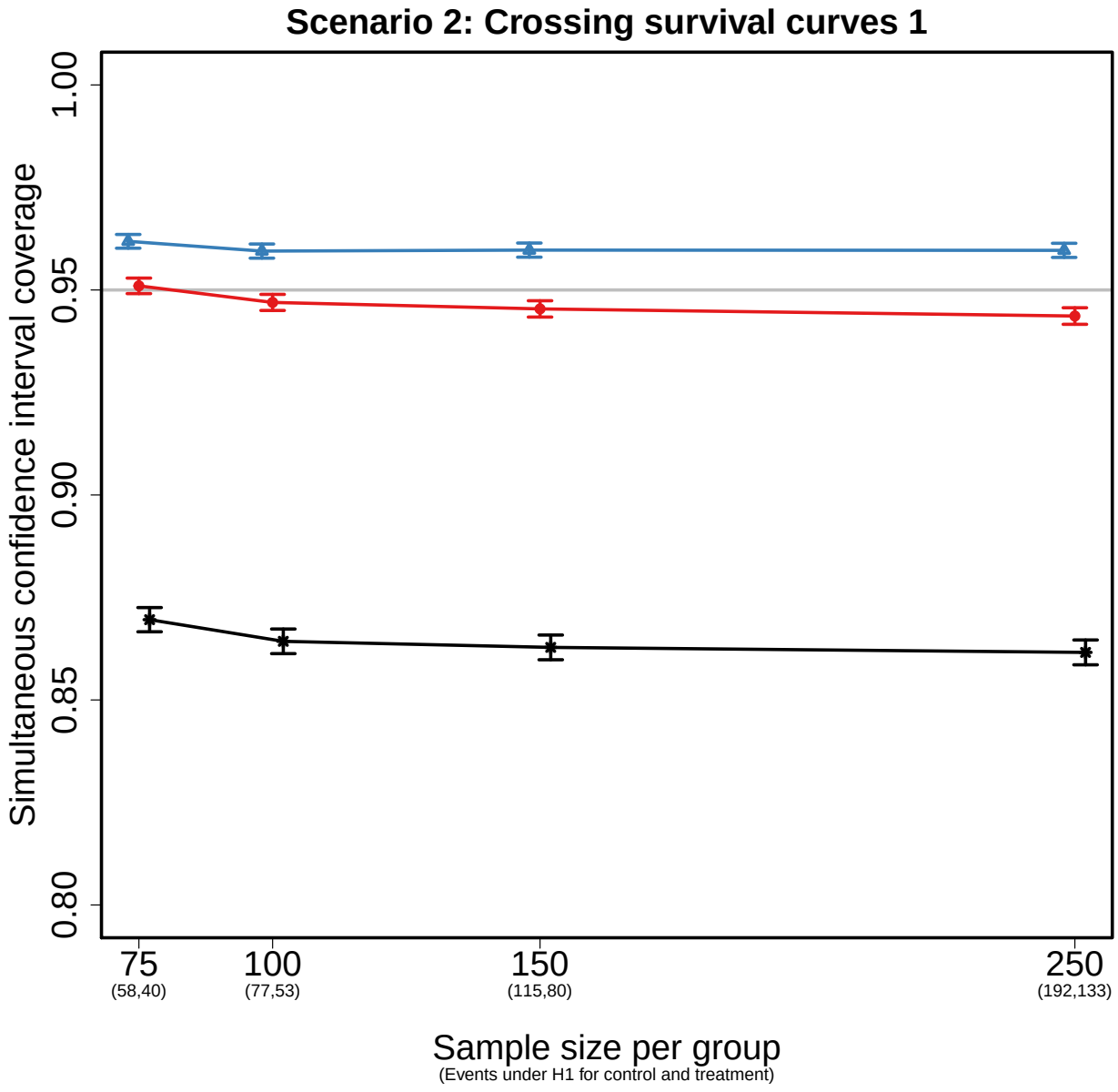
FWER and power to reject at least one H_j

Multivariate normal test (asymptotic Σ), Bonferroni test, Logrank test.



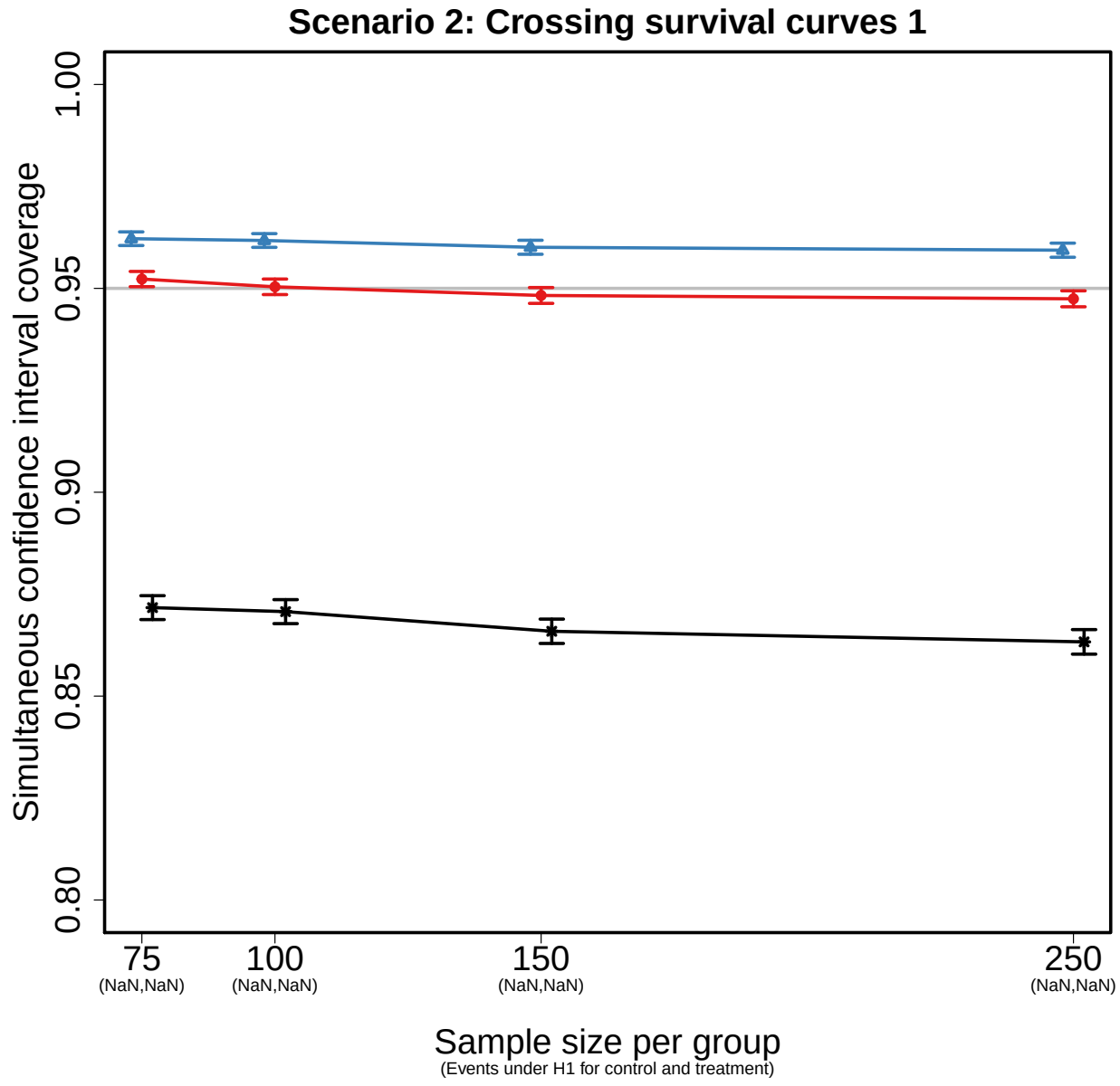
Simultaneous Coverage Probability

Multivariate normal (asymptotic Σ), Bonferroni, Unadjusted.



Simultaneous Coverage Probability (Perturbation)

Multivariate normal (perturbation approach), Bonferroni, Unadjusted.



Properties of the multiple testing procedure

- If quantiles (as median survival) are tested, the perturbation approach has better CI coverage and type I error rate control. In general, better characteristics for log-transformed parameters.
- Power depends on the specific scenario. For crossing survival curves, the logrank test testing multiple parameters can have larger power than the logrank test.
- Tests based on the multivariate normal approximation are more powerful than Bonferroni adjustment.
- The tests will soon be available in the R-package nph `https://cran.r-project.org/web/packages/nph/`

Discussion

- There is no general purpose measure for differences in survival distributions, especially if survival curves cross.
- Therefore, inference procedures addressing several parameters simultaneously can be useful.
- In many settings, this comes at a loss in power (compared to the logrank test) but provides additional information.

Literature I

[1] O. O. Aalen, P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding. History of applications of martingales in survival analysis. *arXiv preprint arXiv:1003.0188*, 2010.

[2] J. W. Bartlett, T. P. Morris, M. J. Stensrud, R. M. Daniel, S. K. Vansteelandt, and C.-F. Burman. The hazards of period specific and weighted hazard ratios. *Statistics in Biopharmaceutical Research*, 12(4):518, 2020.

[3] M. Brückner and W. Brannath. Sequential tests for non-proportional hazards data. *Lifetime data analysis*, 23(3):339–352, 2017.

[4] C. Chauvel and J. O'quigley. Tests for comparing estimated survival functions. *Biometrika*, 101(3):535–552, 2014.

[5] P. Flandre and J. O'Quigley. Comparing kaplan–meier curves with delayed treatment effects: applications in immunotherapy trials. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(4):915–939, 2019.

[6] T. R. Fleming and D. P. Harrington. *Counting processes and survival analysis*. John Wiley & Sons, 1991.

[7] B. Freidlin and E. L. Korn. Methods for accommodating nonproportional hazards in clinical trials: ready for the primary analysis? *Journal of Clinical Oncology*, 37(35):3455, 2019.

[8] J. Irwin. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Epidemiology & Infection*, 47(2):188–189, 1949.

[9] J. D. Kalbfleisch and R. L. Prentice. Estimation of the average hazard ratio. *Biometrika*, 68(1):105–112, 1981.

[10] T. G. Karrison et al. Versatile tests for comparing survival curves based on weighted log-rank statistics. *Stata Journal*, 16(3):678–690, 2016.

[11] H. T. Kim and R. Gray. Three-component cure rate model for nonproportional hazards alternative in the design of randomized clinical trials. *Clinical Trials*, 9(2):155–163, 2012.

[12] E. L. Korn and B. Freidlin. Interim futility monitoring assessing immune therapies with a potentially delayed treatment effect. *Journal of Clinical Oncology*, 36(23):2444–2449, 2018.

[13] H. Li, D. Han, Y. Hou, H. Chen, and Z. Chen. Statistical inference methods for two crossing survival curves: a comparison of methods. *PLoS One*, 10(1), 2015.

[14] D. Magirr and C.-F. Burman. Modestly weighted logrank tests. *Statistics in medicine*, 38(20):3782–3790, 2019.

[15] Y. Park and L. Wei. Estimating subject-specific survival functions under the accelerated failure time model. *Biometrika*, 90(3):717–723, 2003.

Literature II

[16] M. Parzen, L. Wei, and Z. Ying. Simultaneous confidence intervals for the difference of two survival functions. *Scandinavian Journal of Statistics*, 24(3):309–314, 1997.

[17] G. Rauch, W. Brannath, M. Brueckner, and M. Kieser. The average hazard ratio–a good effect measure for time-to-event endpoints when the proportional hazard assumption is violated? *Methods of information in medicine*, 57(03):089–100, 2018.

[18] R. Ristl, N. Ballarini, H. Götte, A. Schüler, M. Posch, and F. König. Delayed treatment effects, treatment switching and heterogeneous patient populations: How to design and analyze rcts in oncology. *Pharmaceutical Statistics*, 2020.

[19] P. Royston. A combined test for a generalized treatment effect in clinical trials with a time-to-event outcome. *The Stata Journal*, 17(2):405–421, 2017.

[20] P. Royston, B. Choodari-Oskooei, M. K. Parmar, and J. K. Rogers. Combined test versus logrank/cox test in 50 randomised trials. *Trials*, 20(1):1–10, 2019.

[21] P. Royston and M. K. Parmar. 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*, 16(1):1–13, 2016.

[22] M. Schumacher. Two-sample tests of cramér–von mises-and kolmogorov–smirnov-type for randomly censored data. *International Statistical Review/Revue Internationale de Statistique*, pages 263–281, 1984.

[23] R. E. Tarone. On the distribution of the maximum of the logrank statistic and the modified wilcoxon statistic. *Biometrics*, pages 79–85, 1981.

[24] L. Zhao, B. Claggett, L. Tian, H. Uno, M. A. Pfeffer, S. D. Solomon, L. Trippa, and L. Wei. On the restricted mean survival time curve in survival analysis. *Biometrics*, 72(1):215–221, 2016.