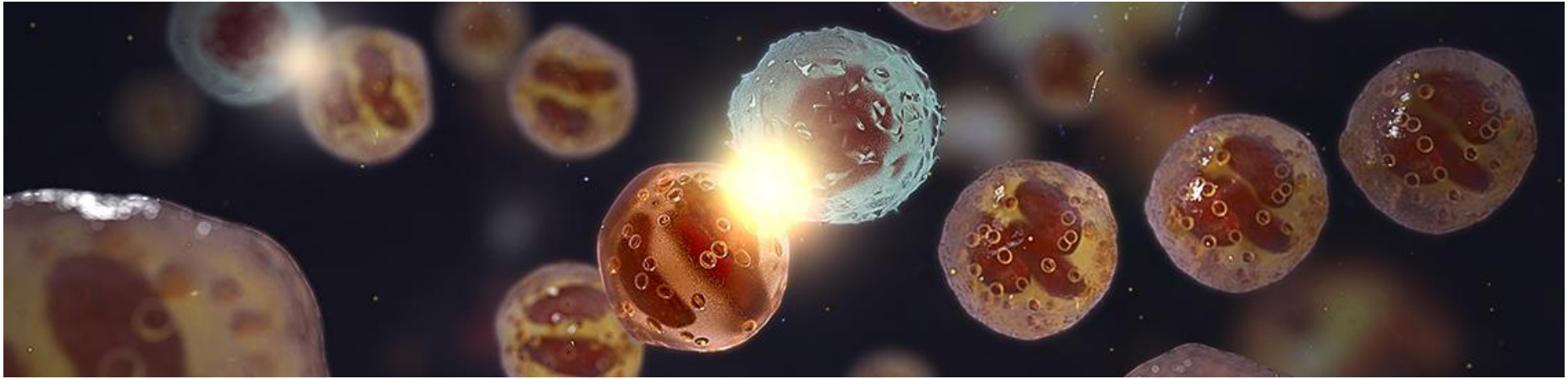
Group-sequential and adaptive designs in immuno-oncology

Dominic Magirr

PSI Conference 2018

June 2018

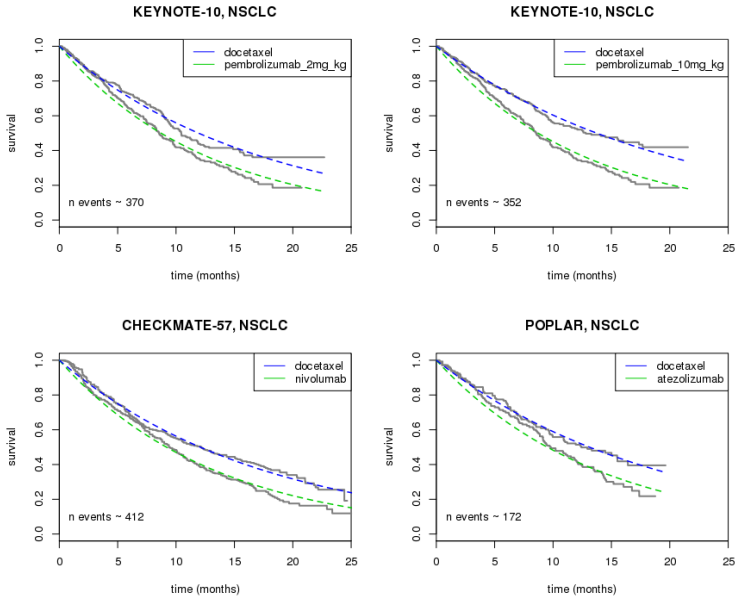


Outline

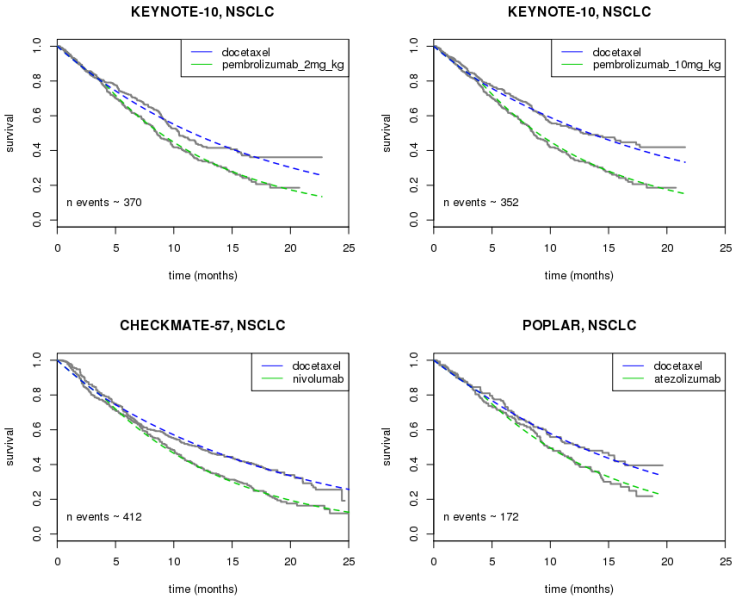
- **Delayed separation of survival curves**
- Futility Analysis
- Efficacy Analysis
- Alternatives to the logrank test / Cox model

Evidence for delayed separation of survival curves (OS)

Exponential model



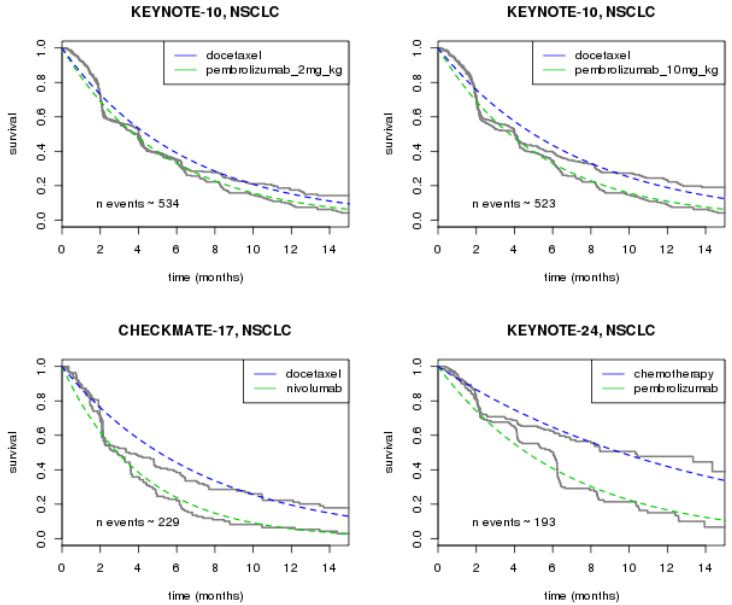
Piecewise exponential model



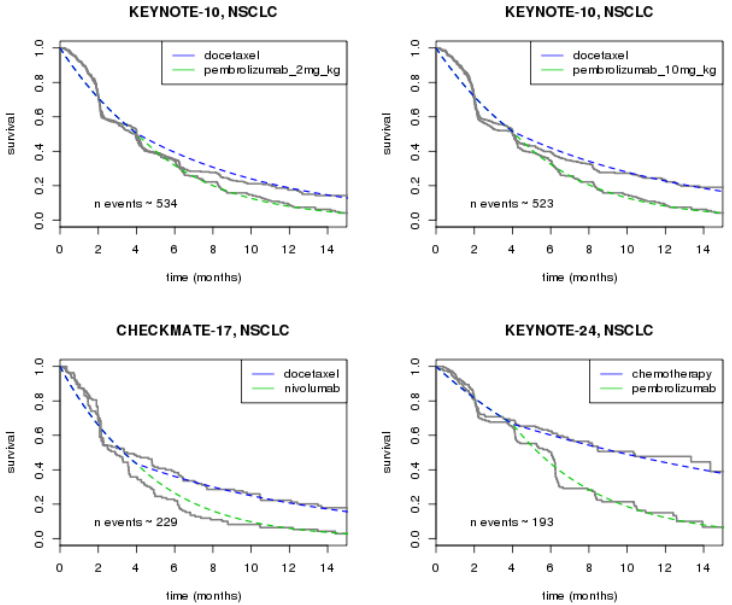
Based on reconstructed data from Kaplan-Meier curves. Herbst (2016); Borghaei (2015); Fehrenbacher (2016).

Evidence for delayed separation of survival curves (PFS)

Exponential model



Piecewise exponential model



Based on reconstructed data from Kaplan-Meier curves. Herbst (2016); Brahmer (2015); Reck (2016).

Outline

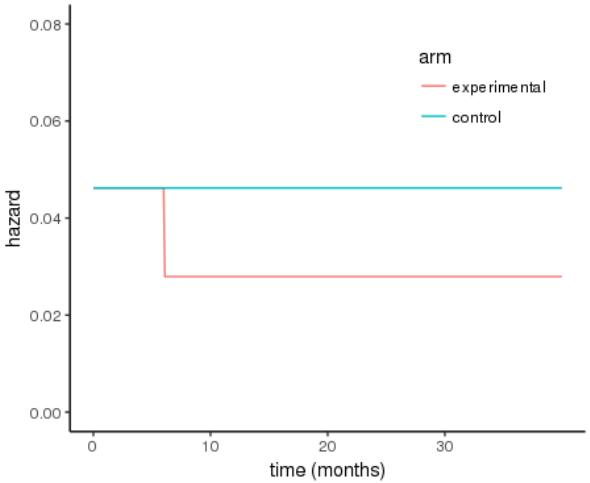
- Delayed separation of survival curves
- **Futility Analysis**
- Efficacy Analysis
- Alternatives to the logrank test / Cox model

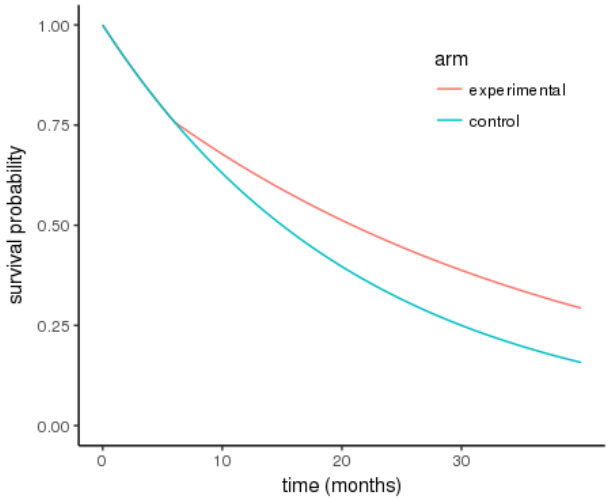
Typical trial design question

- NSCLC.
- “Immuno-oncology agent (IO) + Standard Care (SoC)” vs SoC.
- Details:
  - N = 550 recruited over 18 months.
  - Median PFS on SoC ~15 months.
  - Alternative hypothesis: hazard ratio = 0.7.
  - Final analysis after ~3 years.
- Potential futility analysis at ~ 2 years:
  - Stop trial if, e.g., observed HR > 0.9.
  - Under proportional hazards (HR = 0.7), the risk of stopping is ~3%.
- **What is the impact on operating characteristics if there is a delayed separation in PFS?**

What is risk of stopping at interim when…

..delay = 3m, 4m, 6m, etc., and “average HR” = 0.7.





No unique definition for the average hazard ratio (see Schemper, 2009).

I’m using:

$$\text{avHR} = \exp\left(\int_0^{40} \log\left\{\frac{h_E(t)}{h_C(t)}\right\} f(t)\, dt\right)$$

*Schemper, Michael, et al. "The estimation of average hazard ratios by weighted Cox regression." Statistics in medicine 28.19 (2009): 2473-2489.

Distribution of the logrank statistic

$$U := \sum_i (O_{1,i} - E_{1,i})$$

$$= \sum_{i:t_i<6} (O_{1,i} - E_{1,i}) \quad + \quad \sum_{i:t_i>6} (O_{1,i} - E_{1,i})$$

$$\approx N(\theta_1 V_1, V_1) \qquad \approx N(\theta_2 V_2, V_2)$$

log hazard ratio on (0, 6)

log hazard ratio on (6, ∞)

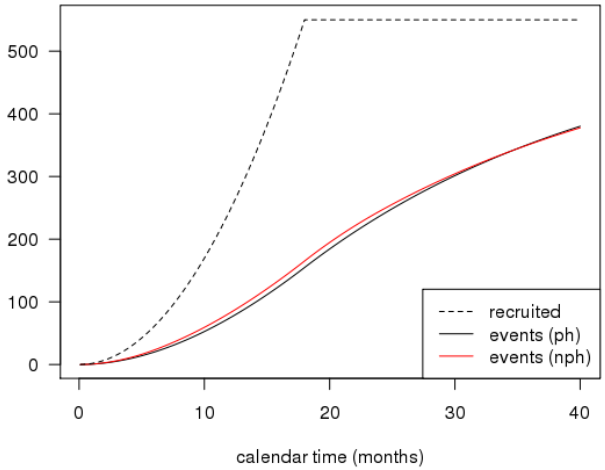
The Peto estimate (Yusuf, 1985) of the log hazard ratio (under proportional hazards) is U / V.

Here,

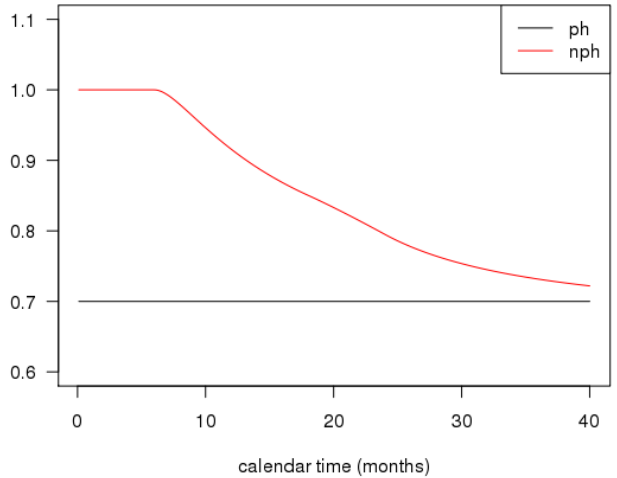
$$E(\log(\widehat{HR})) = E(U / V) \approx \{\theta_1 V_1 + \theta_2 V_2\} / \{V_1 + V_2\}$$

E(U/V) over calendar time

n



exp{E(U/V)}



Probability of stopping for futility

Interim analysis after approximately 24 months / 245 events.

Delay	HR1	HR2	AvHR	exp(E(U/V)) at interim	Pr(Stop) Bound = 0.9	Pr(stop) Bound = 1.05
NA	1	1	1	1	0.81	0.35
NA	0.7	0.7	0.7	0.7	0.03	<0.01
3	1	0.66	0.7	0.74	0.07	<0.01
6	1	0.6	0.7	0.8	0.17	0.02

Idea: use evidence of delayed effect to improve design

1.

Hazard ratio months 0 to 6

Hazard ratio months 6 to 24

$$\hat{\mathrm{HR}} = \exp\left(p_1 \log(\hat{\mathrm{HR}}_1) + p_2 \log(\hat{\mathrm{HR}}_2)\right)$$

Proportion of events from months 0 to 6

Proportion of events from months 6 to 24

2.

Hazard ratio ratio

$$\widehat{HRR} = \widehat{HR}_2 \,/\, \widehat{HR}_1$$

3.

Stop when $\widehat{HR} > 1.05$ if $\widehat{HRR} < 0.8$

$\widehat{HR} > 0.9$ if $\widehat{HRR} \geq 0.8$

Less aggressive stopping boundary when there is evidence of a delayed effect.

Probability of stopping for futility

Interim analysis after approximately 24 months / 245 events

Delay	HR1	HR2	AvHR	exp(E(U/V)) at interim	Pr(Stop) Bound = 0.9	Pr(stop) Bound = 1.05	Pr(stop) Bound based on HRR
NA	1	1	1	1	0.81	0.35	0.72
NA	0.7	0.7	0.7	0.7	0.03	<0.01	0.02
3	1	0.66	0.7	0.74	0.07	<0.01	0.03
6	1	0.6	0.7	0.8	0.17	0.02	0.04

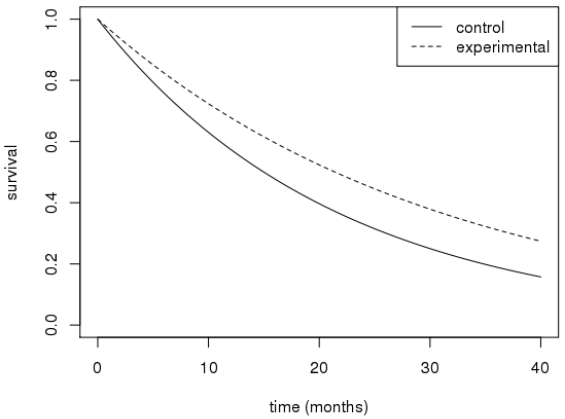
Outline

- Delayed separation of survival curves
- Futility Analysis
- **Efficacy Analysis**
- Alternatives to the logrank test / Cox model

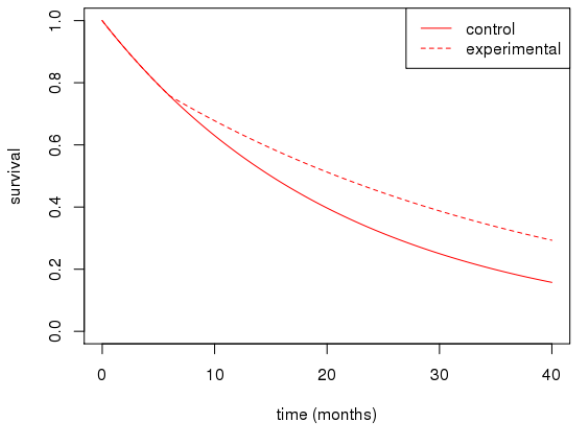
Design dilemma

We want high power for both scenarios…

Proportional hazards (ph)

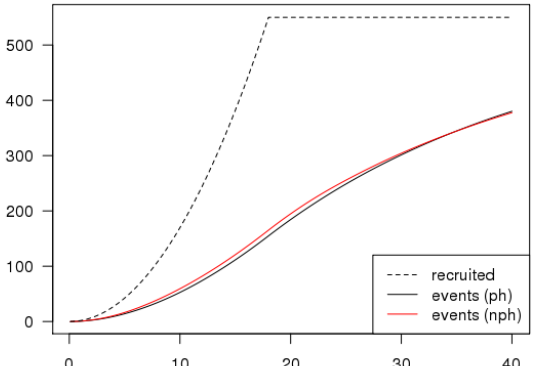


Non proportional hazards (nph)

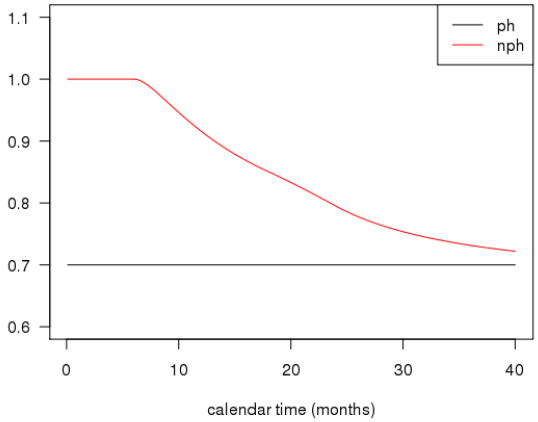


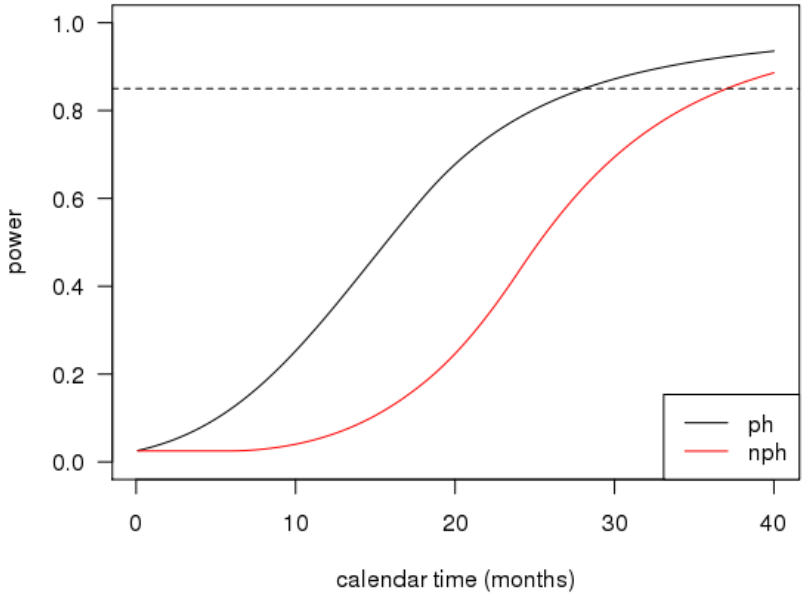
Requires different duration and # events

n

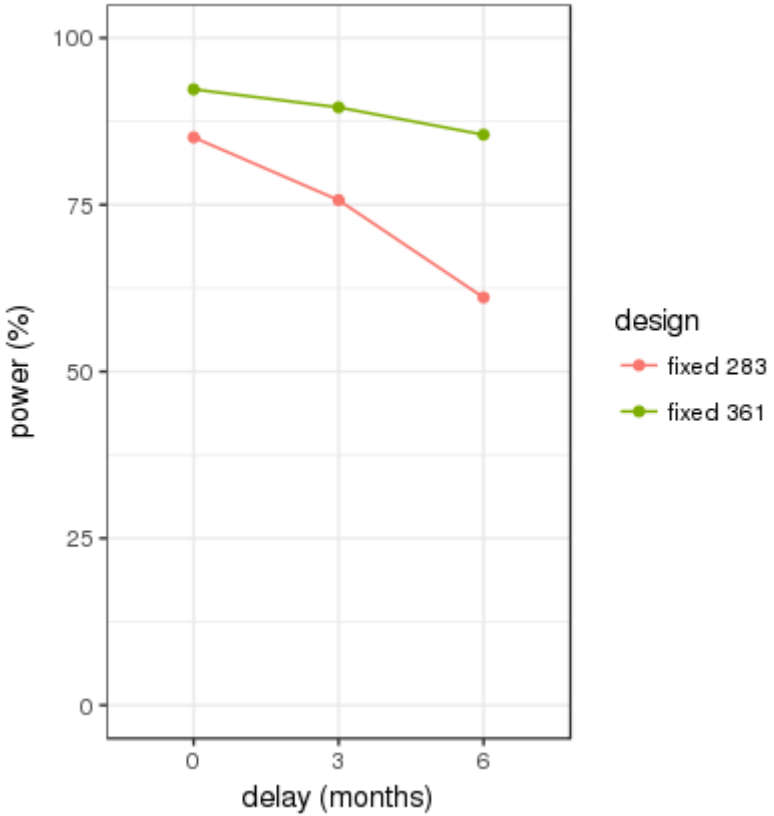


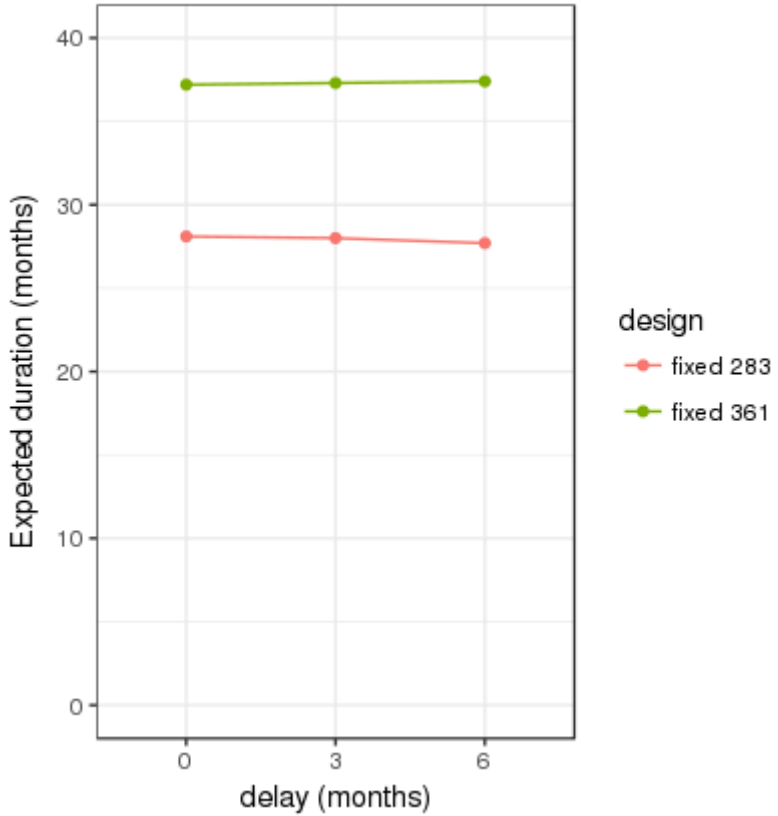
exp{E(U/V)}





Compare designs





Adaptive design

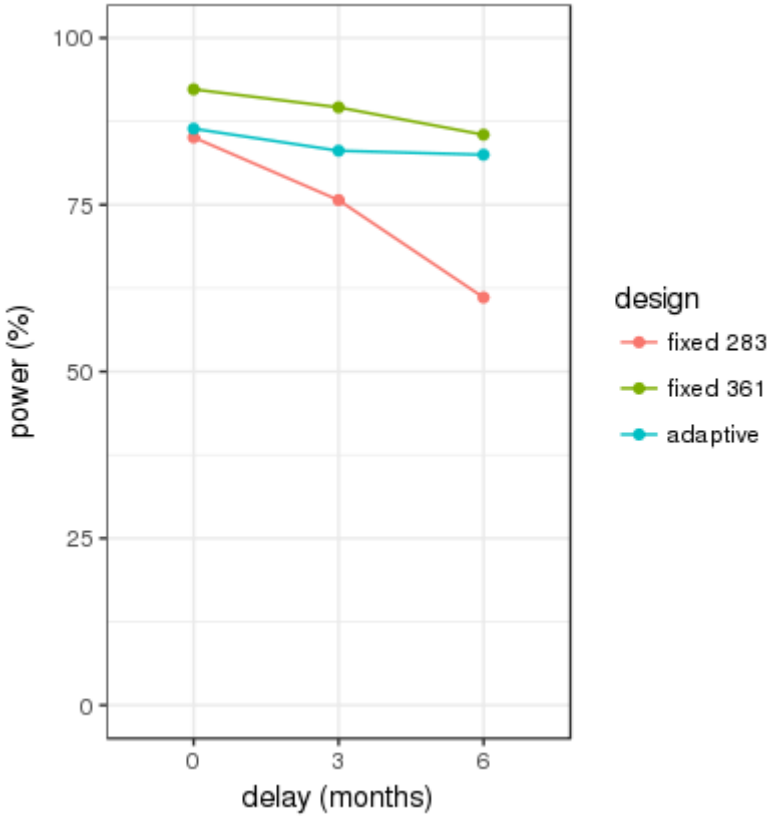
Recall $\widehat{HRR} = \widehat{HR}_2 / \widehat{HR}_1$ and $\hat{\mathrm{HR}} = \exp\left(p_1 \log(\hat{\mathrm{HR}}_1) + p_2 \log(\hat{\mathrm{HR}}_2)\right)$

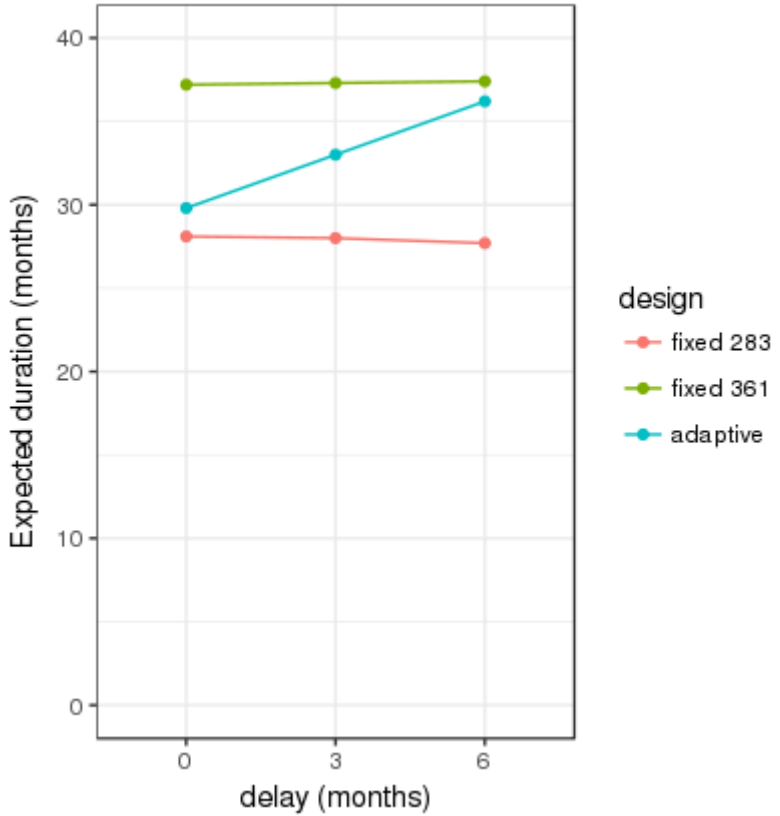
$$\begin{aligned}
\operatorname{cov}\left\{\log(\hat{\mathrm{HR}}_2) - \log(\hat{\mathrm{HR}}_1),\ p_1 \log(\hat{\mathrm{HR}}_1) + p_2 \log(\hat{\mathrm{HR}}_2)\right\} &= p_2 \operatorname{var}\left\{\log(\hat{\mathrm{HR}}_2)\right\} - p_1 \operatorname{var}\left\{\log(\hat{\mathrm{HR}}_1)\right\} \\
&\approx \frac{V_2}{V_1 + V_2}\frac{1}{V_2} - \frac{V_1}{V_1 + V_2}\frac{1}{V_1} \\
&= 0
\end{aligned}$$

- After 283 events, calculate $\widehat{HRR}$.
- If $\widehat{HRR} \geq 0.8$, analyse the trial now.
- If $\widehat{HRR} < 0.8$, continue follow-up to 361 events.
- Apply standard analysis techniques. Test as normal at 0.025 level.

Similar ideas in Qiu, Peihua, and Jun Sheng. "A two-stage procedure for comparing hazard rate functions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70, no. 1 (2008): 191-208.

Compare designs





But what about a group-sequential design?

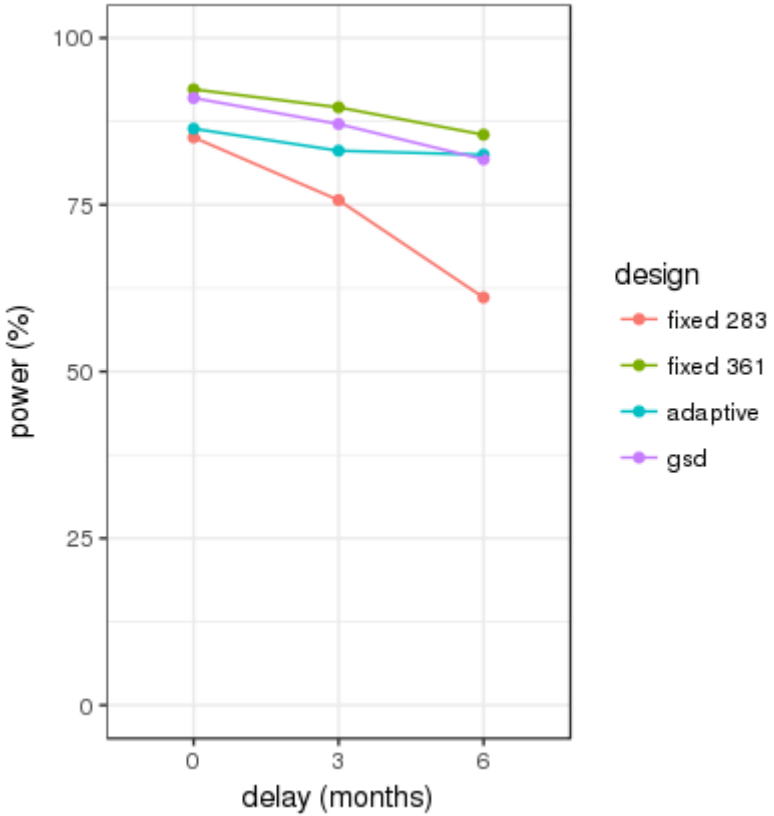
Adaptive design

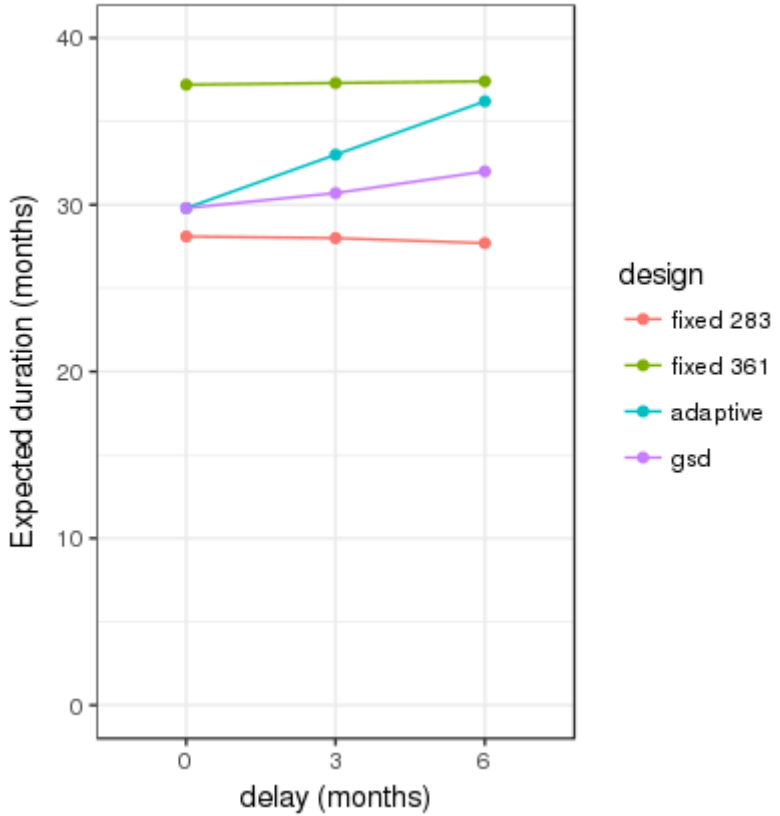
- Analysis at either 283 or 361 events.
- No alpha penalty.
- Reject H_0 if $Z > 1.96$.

Group sequential design

- First analysis at 283 events.
- Reject H_0 if $Z_1 > 2.11$.
- Otherwise, continue to 361 events.
- Reject H_0 if $Z_2 > 2.12$.

Compare designs





Outline

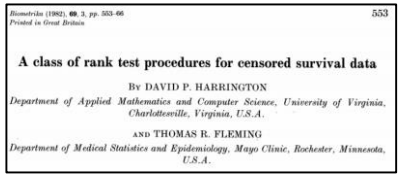
- Delayed separation of survival curves
- Futility Analysis
- Efficacy Analysis
- **Alternatives to the logrank test / Cox model**

Alternatives to the logrank test

Logrank test $U := \sum_i (O_{1,i} - E_{1,i})$

- Most powerful non-parametric test under proportional hazards.
- No reason we have to stick with it when we expect non-proportional hazards.

Weighted logrank tests


Biometrika (1982), 69, 3, pp. 553–66
Printed in Great Britain

553

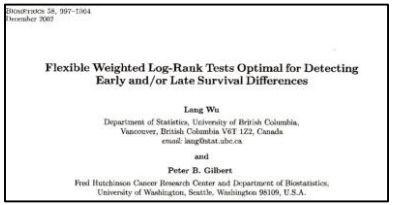
A class of rank test procedures for censored survival data

By DAVID P. HARRINGTON

Department of Applied Mathematics and Computer Science, University of Virginia, Charlottesville, Virginia, U.S.A.

AND THOMAS R. FLEMING

Department of Medical Statistics and Epidemiology, Mayo Clinic, Rochester, Minnesota, U.S.A.

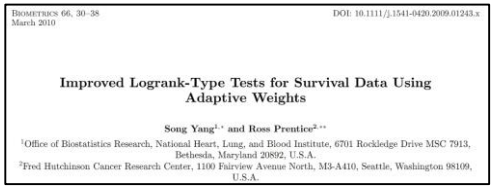

BIOMETRICS 58, 997–1004
December 2002

Flexible Weighted Log-Rank Tests Optimal for Detecting Early and/or Late Survival Differences

Lang Wu
Department of Statistics, University of British Columbia, Vancouver, British Columbia V6T 1Z2, Canada
email: lang@stat.ubc.ca

and

Peter B. Gilbert
Fred Hutchinson Cancer Research Center and Department of Biostatistics, University of Washington, Seattle, Washington 98109, U.S.A.


BIOMETRICS 66, 30–38
March 2010

DOI: 10.1111/j.1541-0420.2009.01243.x

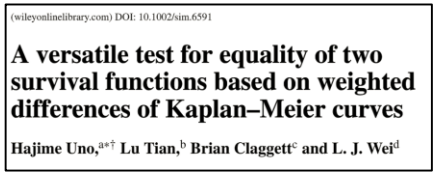
Improved Logrank-Type Tests for Survival Data Using Adaptive Weights

Song Yang[1,*] and Ross Prentice[2,**]

[1]Office of Biostatistics Research, National Heart, Lung, and Blood Institute, 6701 Rockledge Drive MSC 7913, Bethesda, Maryland 20892, U.S.A.

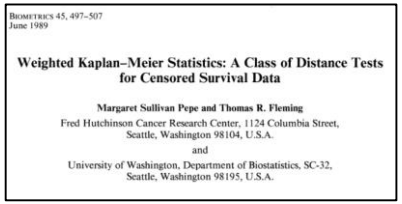
[2]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, M3-A410, Seattle, Washington 98109, U.S.A.

Weighted Kaplan-Meier-based tests


(wileyonlinelibrary.com) DOI: 10.1002/sim.6591

A versatile test for equality of two survival functions based on weighted differences of Kaplan–Meier curves

Hajime Uno,[a*†] Lu Tian,[b] Brian Claggett[c] and L. J. Wei[d]


BIOMETRICS 45, 497–507
June 1989

Weighted Kaplan–Meier Statistics: A Class of Distance Tests for Censored Survival Data

Margaret Sullivan Pepe and Thomas R. Fleming
Fred Hutchinson Cancer Research Center, 1124 Columbia Street, Seattle, Washington 98104, U.S.A.
and
University of Washington, Department of Biostatistics, SC-32, Seattle, Washington 98195, U.S.A.


Royston and Parmar *BMC Medical Research Methodology* 2013, **13**:152
http://www.biomedcentral.com/1471-2288/13/152

BMC Medical Research Methodology

RESEARCH ARTICLE Open Access

Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome

Patrick Royston* and Mahesh KB Parmar

Logrank vs weighted logrank

$$U_W = \sum_i w_i (O_{1i} - E_{1i})$$

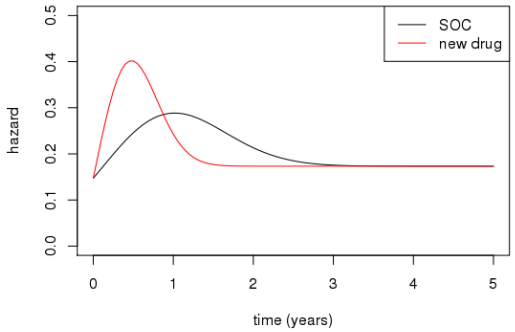
Where:

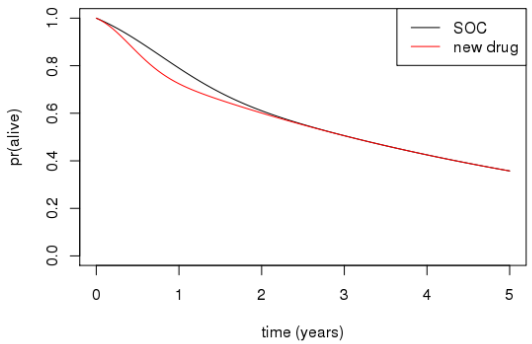
- Events occur at times $t_1 < t_2 < \cdots < t_n$
- O_{1i} is the observed number of events on the control arm at time t_i
- E_{1i} is the expected number of events on the control arm at time t_i, under the point null hypothesis.

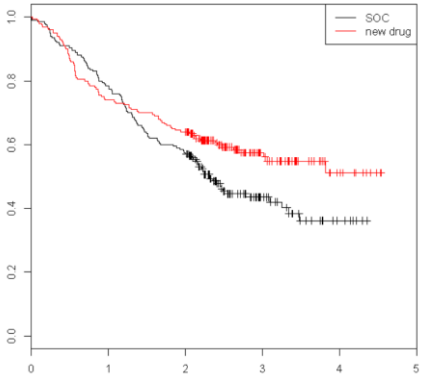
Hopefully, we will see "O_{1i} = 1" more often than "O_{1i} = 0". But we think this is more likely at later timepoints so we will let w_i increase with time. In this case U_W >> 0, the corresponding p-value will be low, and we reject then null.

Logrank vs weighted logrank

The problem is that survival can be uniformly worse on the new drug, but because the hazards cross, we end up with a statistically significant “benefit” more than 2.5% (say) of the time…



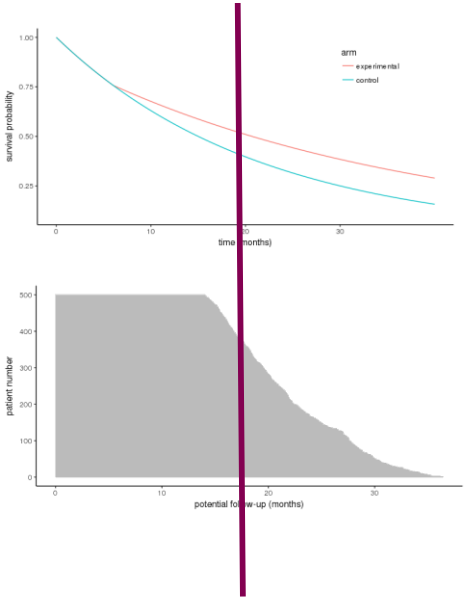




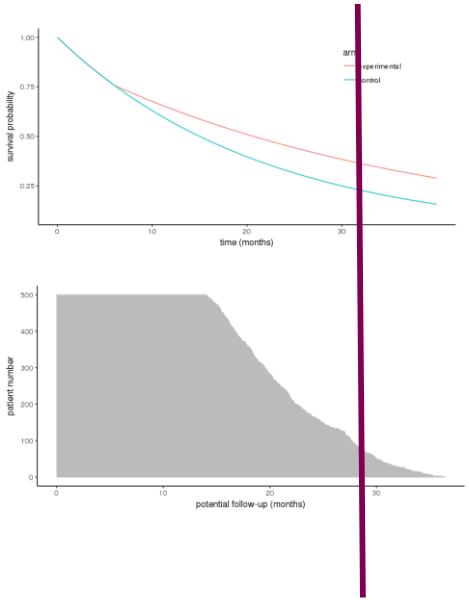
- Requires careful weighting.
- Work in progress:
  - Burman, Magirr, Bartlett (2018) “One-sided weighted log-rank tests can be severely misleading”. *In preparation.*
  - Magirr, Bartlett, Burman (2018) “Modestly weighted log-rank tests”. *In preparation.*

Logrank vs Kaplan-Meier based tests

Either we ignore a large portion of the data…



..or we blow up the variance.



Requires careful weighting (see Pepe & Fleming, 1989)

Summary

Futility analysis

- Extra caution required when stopping for futility.
- Worth exploring non-standard stopping rules, maybe HRR or looking very carefully at the KM curves.

Efficacy analysis

- Group sequential designs work very well. The flexibility they offer (in terms of spending functions) lets us search for a specific design with robust power over the scenarios we are most interested in.

Alternatives to logrank / Cox model

- No reason to always use the logrank / Cox model when we expect non-proportional hazards.
- Alternative methods out there.
- We still need a better understanding of their type I error and power.

References

Herbst, Roy S., et al. "Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial." *The Lancet* 387.10027 (2016): 1540-1550.

Borghaei, Hossein, et al. "Nivolumab versus docetaxel in advanced nonsquamous non–small-cell lung cancer." *New England Journal of Medicine* 373.17 (2015): 1627-1639.

Fehrenbacher, Louis, et al. "Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial." *The Lancet* 387.10030 (2016): 1837-1846.

Brahmer, Julie, et al. "Nivolumab versus docetaxel in advanced squamous-cell non–small-cell lung cancer." *New England Journal of Medicine* 373.2 (2015): 123-135.

Reck, Martin, et al. "Pembrolizumab versus chemotherapy for PD-L1–positive non–small-cell lung cancer." *New England Journal of Medicine* 375.19 (2016): 1823-1833.

Schemper, Michael, Samo Wakounig, and Georg Heinze. "The estimation of average hazard ratios by weighted Cox regression." *Statistics in medicine* 28.19 (2009): 2473-2489.

Yusuf, Salim, et al. "Beta blockade during and after myocardial infarction: an overview of the randomized trials." *Progress in cardiovascular diseases* 27.5 (1985): 335-371.

Qiu, Peihua, and Jun Sheng. "A two-stage procedure for comparing hazard rate functions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1 (2008): 191-208.

Royston, Patrick, and Mahesh KB Parmar. "Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome." *BMC medical research methodology* 13.1 (2013): 152.

Pepe, Margaret Sullivan, and Thomas R. Fleming. "Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data." *Biometrics* (1989): 497-507.

Uno, Hajime, et al. "A versatile test for equality of two survival functions based on weighted differences of Kaplan–Meier curves." *Statistics in medicine* 34.28 (2015): 3680-3695.

Wu, Lang, and Peter B. Gilbert. "Flexible Weighted Log-Rank Tests Optimal for Detecting Early and/or Late Survival Differences." *Biometrics* 58.4 (2002): 997-1004.

Harrington, David P., and Thomas R. Fleming. "A class of rank test procedures for censored survival data." *Biometrika* 69.3 (1982): 553-566.

Yang, Song, and Ross Prentice. "Improved Logrank-Type Tests for Survival Data Using Adaptive Weights." *Biometrics*66.1 (2010): 30-38.

Confidentiality Notice