Design and Analysis of Clinical Trials in the Presence of Non-Proportional Hazards

Keaven M. Anderson, Merck Research Laboratories Satrajit Rochoudhury, Pfizer, Inc.

September 13, 2018

Disclaimer

While the authors represent their companies as well as the Non-proportional Hazards Working Group, opinions should be considered their own.

Overview

- History
- Summary of methods and results
- Challenges
  - Where does proposed alternative break down?
  - How can we meaningfully describe treatment group differences?
  - Examples
- Where are we now?
- Conclusions

History of the NPH working group

- Problem: Methods for proportional hazards may not be effective generally
- 2016: FDA met with several PhRMA reps to discuss non-proportional hazards (NPH) issues
  - Working group formed including FDA and industry
- 2017: Preliminary work presented in mid-year industry/FDA discussion
- 2018: Duke-Margolis conference with discussion of preliminary discussions
  - https://healthpolicy.duke.edu/events/public-workshop-oncology-clinical-trials-presence-non-proportional-hazards
- Today: update!

Motivation

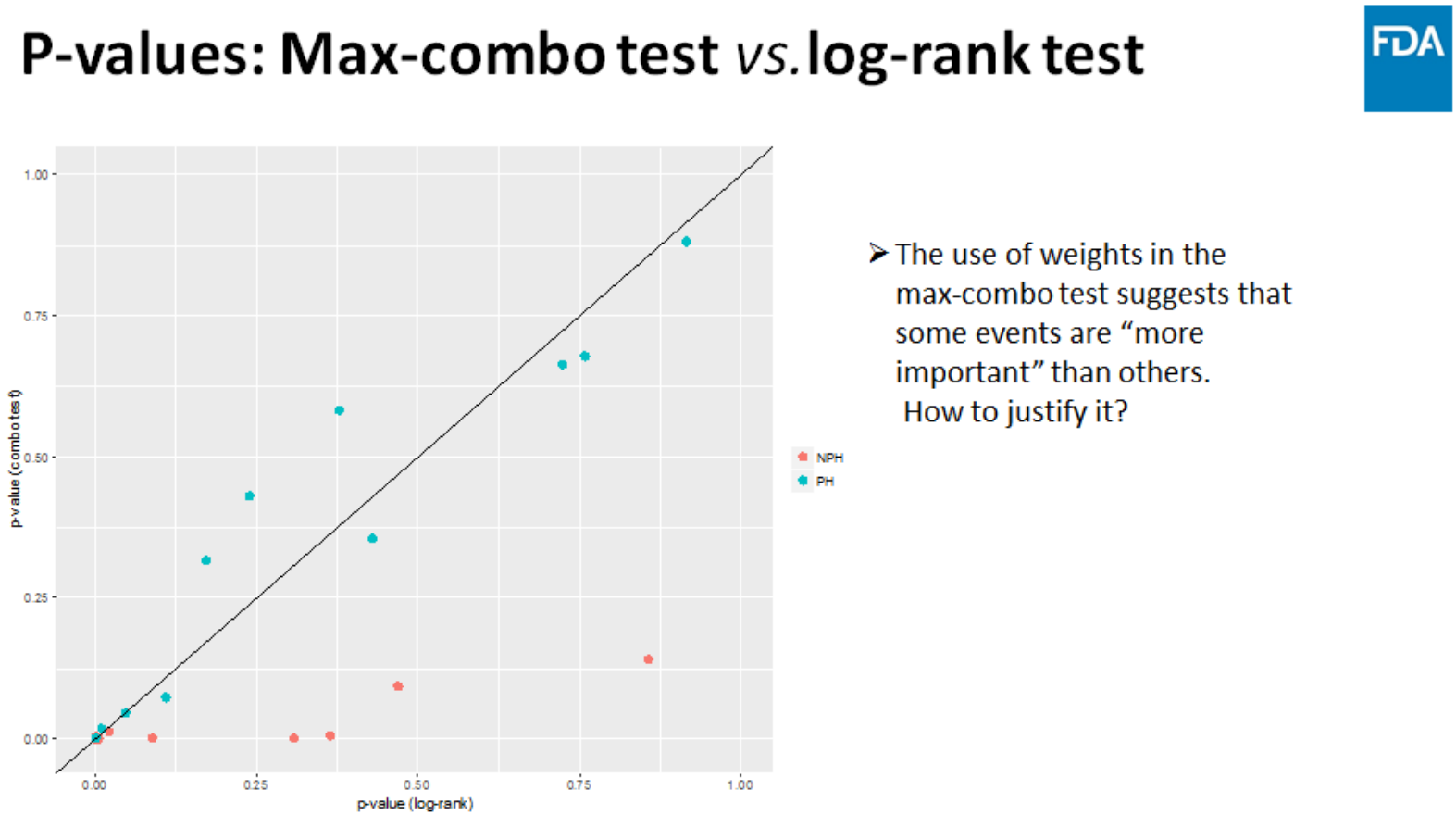


Figure 1: Should some trials not positive by logrank be considered for regulatory approval?

Lijun Zhang, FDA, Duke-Margolis conference, 2018

Methods considered here

- Logrank/Cox HR/Medians (Traditional)
- MaxCombo/Weighted HR (Robust power)
  - Maximum of logrank and 3 Fleming-Harrington weighted logrank tests
  - FH(0,1) - downweight early to detect late effect
  - FH(1,0) - downweight late to detect early effect
  - FH(1,1) - downweight early and late to detect middle effect
  - Correlation known to adjust p-value (Karrison and others (2016))
- MaxCombo/Weighted HR - modified
  - e.g., require upper CI for Cox HR is < 1.1
- RMST, RMTL (Complementary estimand and test; Royston and Parmar (2011), Uno et al. (2014))
  - Asymptotics now justified to use late cutoff (LJ Wei, Lu Tian)
- % favorable (Buyse (2010), Péron et al. (2016), Péron et al. (2018), generalized pairwise comparison)
- Weighted Kaplan-Meier Pepe and Fleming (1989), Pepe and Fleming (1991)

Methods and results previously presented

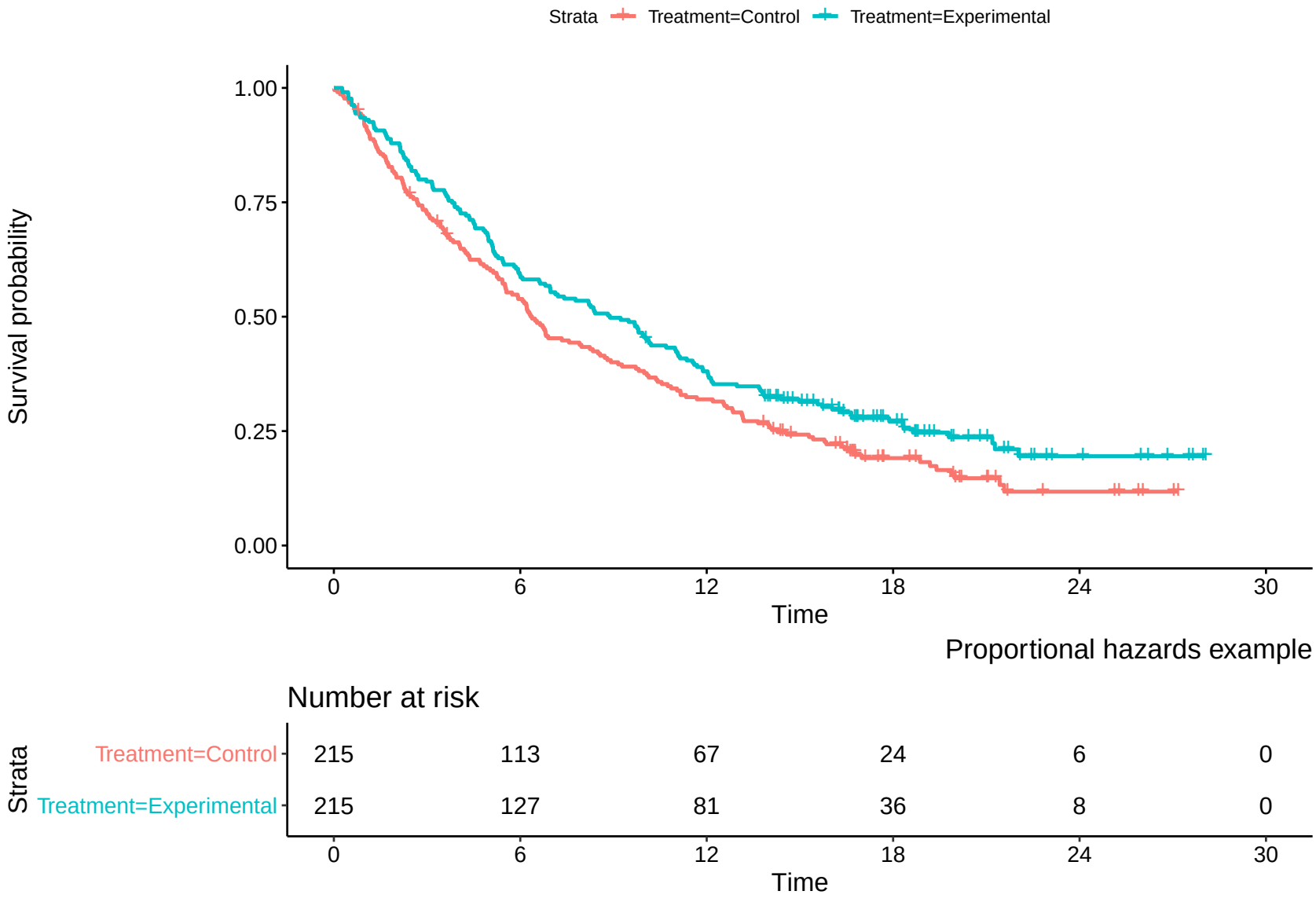
- Duke-Margolis conference
  - Overview of the problem
  - Methods described
  - Design suggestions (get enough follow-up!)
  - Simulation
  - MaxCombo found to be powerful across many scenarios
  - Paper near ready for submission
- Note: Type I error issues for MaxCombo
  - No issues if no treatment difference
  - Potential concern if control benefit decreases over time
- This work assumes subgroups where treatment is more effective are not known

Underlying assumptions, design and examples

- Examples based on simulation similar to extensive experience
- Metastatic disease: 8 month control median
- 15 months enrollment, 30 month study duration
- Design based on HR=0.7
  - Safety and efficacy bounds discussed later
- Potential for non-proportional hazards (examples)
  - Delayed effect: HR=0.5 after 8 months
  - Moderate crossing hazards: HR=1.5 until month 7, HR=0.5 afterwards
  - Severe crossing hazards: HR=1.5 until month 8, HR=0.5 afterwards
    - Cross at 16 months
  - Also include proportional hazards example with HR=0.8

Proportional hazards example

This is an example where logrank is positive and MaxCombo fails



Methods comparison at final analysis

Proportional hazards example

Analysis	Experimental	Control	Estimate	CI	p-value
Median/HR/logrank	8.883	6.378	0.789	(0.637,0.977)	0.015
Weighted HR/MaxCombo	NA	NA	0.789	(0.618,1.008)	0.030
RMST	11.623	9.630	1.993	(0.255,3.732)	0.012
RMTL	15.556	17.549	0.886	(0.797,0.986)	0.013
% favorable	54.784	42.802	11.981	(1.197,23.494)	0.020
Weighted KM	NA	NA	NA	(NA,NA)	0.016

- Estimates are difference for RMST, % favorable; ratio for HR, RMTL
- Positive for logrank, RMST, RMTL, weighted KM.
- Negative for MaxCombo, generalized pairwise comparison.
- This demonstrates the cost of the MaxCombo multiplicity adjustment.
- Generally, power loss is low under proportional hazards
  - Small increase in sample size required to maintain power for MaxCombo

Milestones and piecewise exponential rates

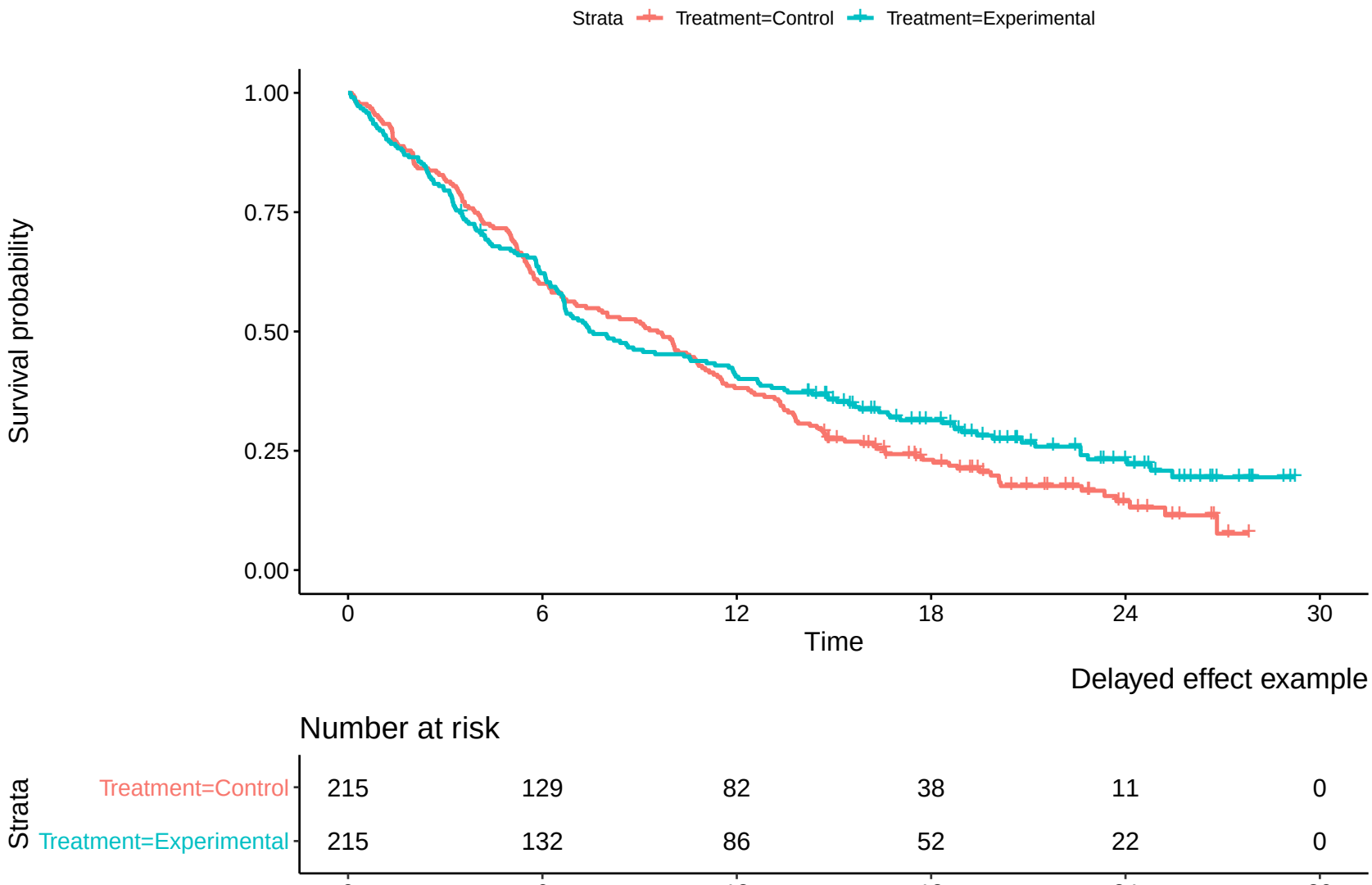
Proportional hazards example

Month	Experimental	Control	Difference	95% CI
3	79.5%	72.4%	7.1%	(-0.9%,15.2%)
6	59.1%	53.9%	5.2%	(-4.2%,14.6%)
12	38.1%	31.9%	6.1%	(-2.9%,15.2%)
18	27.1%	19.1%	8%	(-0.2%,16.2%)
24	19.5%	11.8%	7.8%	(-1%,16.5%)

Period	Experimental	Control	HR	95% CI
0-3 months	0.076	0.106	0.711	(0.482,1.051)
3-6 months	0.098	0.098	0.993	(0.646,1.529)
6-12 months	0.071	0.090	0.788	(0.523,1.189)
>12 months	0.058	0.087	0.661	(0.398,1.098)

Delayed effect example

Delayed benefit example fails by logrank. Only positive test is MaxCombo. The logrank is the closest to significant among others; RMST, RMTL, generalized pairwise all fail.



Methods comparison at final analysis

Delayed effect example

Analysis	Experimental	Control	Estimate	CI	p-value
Median/HR/logrank	7.450	9.557	0.859	(0.693,1.065)	0.083
Weighted HR/MaxCombo	NA	NA	0.736	(0.553,0.979)	0.016
RMST	12.164	11.169	0.995	(-0.799,2.789)	0.138
RMTL	15.644	16.639	0.940	(0.841,1.051)	0.140
% favorable	50.798	47.449	3.349	(-8.377,14.032)	0.286
Weighted KM	NA	NA	NA	(NA,NA)	0.253

- Only MaxCombo positive

Milestones and piecewise exponential rates

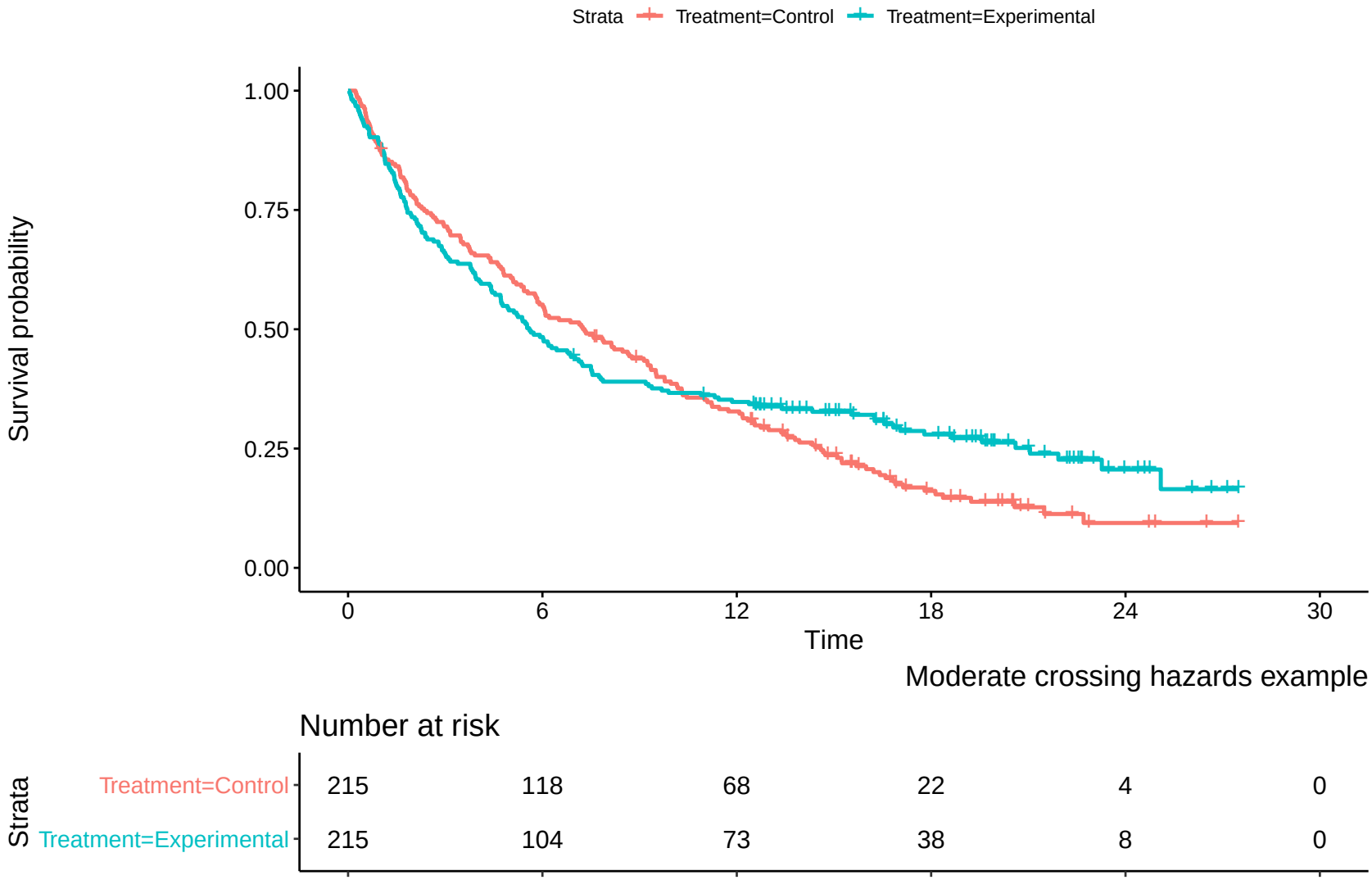
Delayed effect example

Month	Experimental	Control	Difference	95% CI
3	79.5%	81.9%	-2.3%	(-9.8%,5.1%)
6	62.2%	60%	2.2%	(-7%,11.4%)
12	40.5%	38.1%	2.4%	(-6.9%,11.6%)
18	31.4%	23.1%	8.3%	(-0.3%,16.8%)
24	23.2%	14.4%	8.8%	(0.3%,17.2%)

Period	Experimental	Control	HR	95% CI
0-3 months	0.076	0.067	1.143	(0.743,1.759)
3-6 months	0.083	0.102	0.817	(0.531,1.257)
6-12 months	0.075	0.073	1.033	(0.688,1.55)
>12 months	0.047	0.086	0.543	(0.345,0.855)

Moderate crossing hazards example

Arguably tolerable detriment early that might be justified either by the moderate late benefit or detection of an obvious subgroup that can be validated in some way (e.g., KRAS finding).



Methods comparison at final analysis

Moderate crossing example

Analysis	Experimental	Control	Estimate	CI	p-value
Median/HR/logrank	5.594	7.303	0.878	(0.708,1.089)	0.118
Weighted HR/MaxCombo	NA	NA	0.689	(0.515,0.923)	0.004
RMST	10.544	9.503	1.041	(-0.767,2.849)	0.130
RMTL	16.941	17.982	0.942	(0.849,1.046)	0.131
% favorable	48.983	49.432	-0.449	(-11.6,10.82)	0.531
Weighted KM	NA	NA	NA	(NA,NA)	0.367

- Only MaxCombo positive

Milestones and piecewise exponential rates

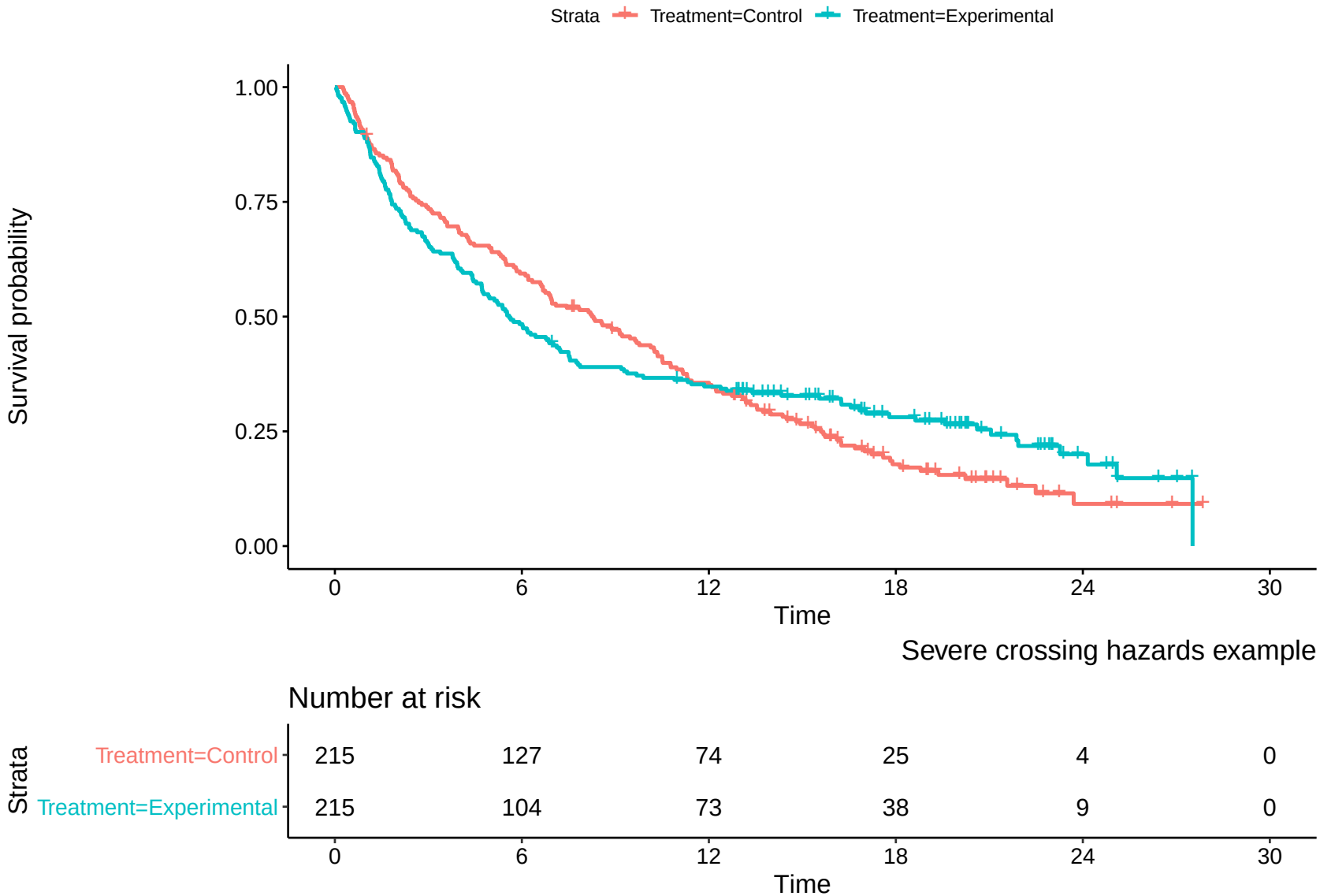
Moderate crossing hazards example

Month	Experimental	Control	Difference	95% CI
3	66%	71.5%	-5.5%	(-14.2%,3.3%)
6	48.4%	55.2%	-6.8%	(-16.2%,2.6%)
12	34.8%	32.8%	2%	(-7%,11%)
18	27.9%	16.1%	11.8%	(3.6%,20.1%)
24	20.6%	9.4%	11.2%	(2.1%,20.3%)

Period	Experimental	Control	HR	95% CI
0-3 months	0.139	0.113	1.237	(0.88,1.737)
3-6 months	0.103	0.086	1.194	(0.754,1.89)
6-12 months	0.058	0.087	0.666	(0.419,1.057)
>12 months	0.038	0.109	0.350	(0.199,0.616)

Severe crossing hazards example

Larger early detriment and less benefit in the tail. Not approvable? Approvable in subset?



Methods comparison at final analysis

Severe crossing example

Analysis	Experimental	Control	Estimate	CI	p-value
Median/HR/logrank	5.594	8.248	0.950	(0.766,1.178)	0.320
Weighted HR/MaxCombo	NA	NA	0.738	(0.551,0.989)	0.019
RMST	10.477	10.140	0.337	(-1.468,2.142)	0.357
RMTL	17.043	17.380	0.981	(0.883,1.089)	0.358
% favorable	47.166	52.834	-5.668	(-17.457,5.883)	0.832
Weighted KM	NA	NA	NA	(NA,NA)	0.667

- Only MaxCombo positive
- MaxCombo becomes negative if upper CI must be < 1.1

Milestones and piecewise exponential rates

Severe crossing hazards example

Month	Experimental	Control	Difference	95% CI
3	66%	73.4%	-7.4%	(-16%,1.3%)
6	48.4%	59.4%	-11%	(-20.4%,-1.6%)
12	34.8%	35.6%	-0.8%	(-9.9%,8.2%)
18	28.1%	17.8%	10.3%	(1.9%,18.7%)
24	20%	9.2%	10.8%	(1.5%,20.1%)

Period	Experimental	Control	HR	95% CI
0-3 months	0.139	0.103	1.349	(0.954,1.908)
3-6 months	0.103	0.070	1.459	(0.904,2.355)
6-12 months	0.058	0.085	0.682	(0.431,1.077)
>12 months	0.043	0.103	0.415	(0.244,0.705)

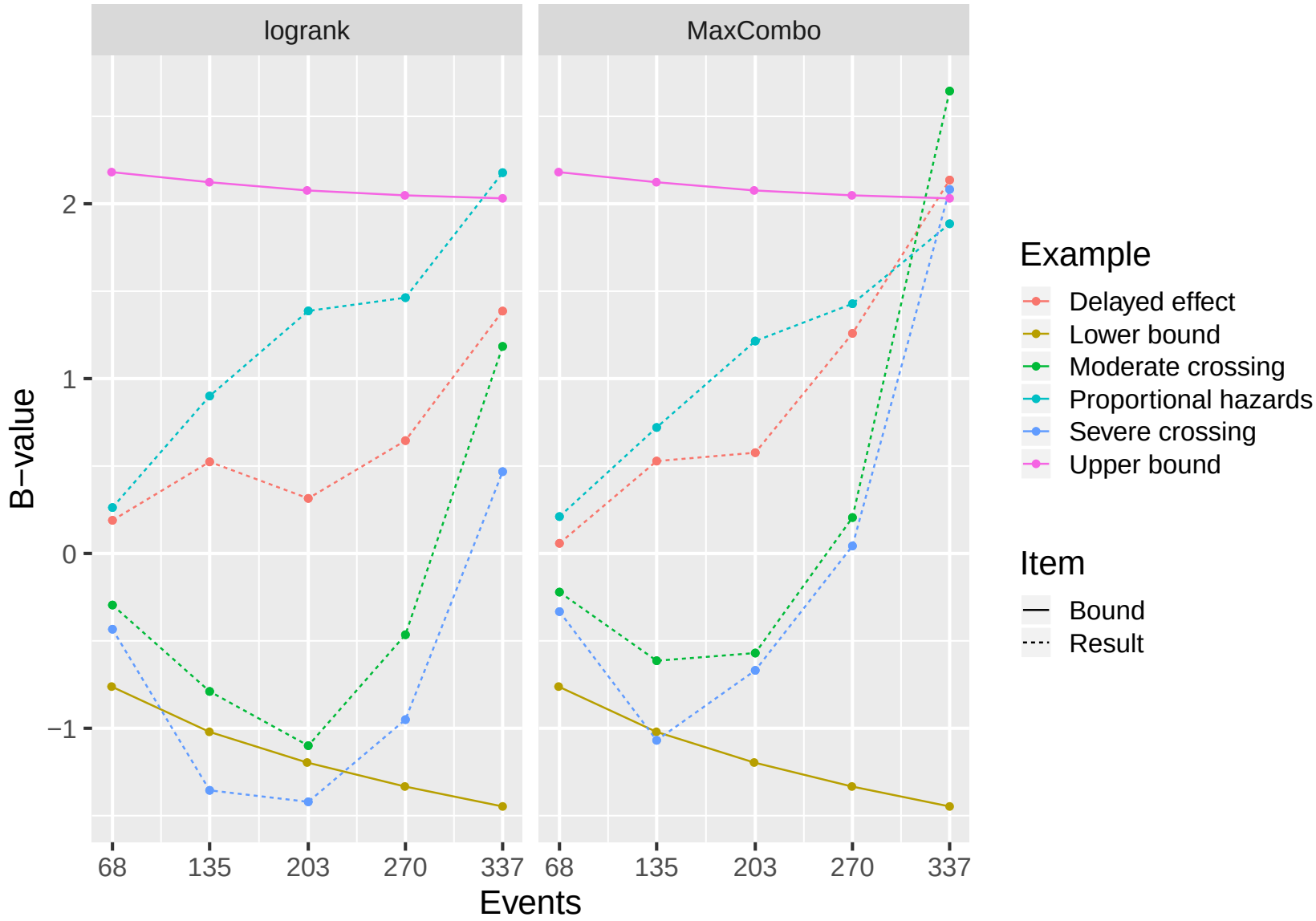
Interim monitoring

- Assume setting where hazards may cross
  - Monitoring for early bad effect can be important
- Interims at 20%, 40%, 60%, 80%
  - First 2 primarily for guidance for bad effects
  - Asymmetric bounds, both with spending under NULL
    - 15% spending for safety; Pocock spending
    - O'Brien-Fleming for efficacy
- B-values on next slide
  - Alternate scale from Z-statistic
  - Brownian process with independent increments
  - Trend is straight line under proportional hazards

Formal group sequential analysis proposal for MaxCombo

- Interim analyses performed with logrank
  - MaxCombo supportive
- Final analysis with MaxCombo
  - Correlation structure understood for multipliticity adjustment
- Sample size for fixed design can be based on Hasegawa proposal
- None of the above REQUIRES simulation; all asymptotics

Interim monitoring



Concerns?

Potential concerns for alternative methods for regulatory approval

- Focus here on metastatic (high-risk) scenario
  - Long-term outcomes with low rates may require alternate approach
- Proposed estimand for MaxCombo not intuitive
  - Weighted HR based on best FH weighting
  - Descriptive alternatives
    - Milestones, piecewise rates and piecewise HR
- Type I error for theoretical cases with no benefit
  - Sponsor needs to justify Type I error protection
  - FURTHER CLARIFICATION NEEDED.
- Primary concern was delayed treatment effect
  - Alternatives other than weighted approaches not doing well?
- Little indication of regulatory interest at this point

Example summary

- Extra protection of Type I error required for crossing hazards
  - Interim safety bounds
  - Confidence bound requirements
- Breakdown issues for MaxCombo now illustrated
  - Are the situations addressed important?
- Descriptive summaries of outcomes useful
  - Complementary estimand approach weakness
- MaxCombo approach positive where others fail
  - Is this good or bad?
  - Are some of these false positives based on underlying scenarios?
  - Is modification requiring Cox HR CI < 1.1 helpful?

Where is NPH Working Group now?

- Near-final draft of simulation paper
- Draft paper on design and analysis prepared
- Estimand working group now working in parallel
- Need for further regulatory interaction

Conclusions

- MaxCombo useful for non-proportional hazards in metastatic setting
- Important benefit could be missed with other methods
- Proposals are ready for alternatives to logrank/Cox/median
- Sponsors encouraged to submit as supportive
- Further discussion needed to move approaches to primary
- Should a trial be consider for Complex Innovative Design Pilot?

References

Buyse, Marc. 2010. "Generalized Pairwise Comparisons of Prioritized Outcomes in the Two-Sample Problem." *Statistics in Medicine* 29 (30). Wiley Online Library: 3245–57.

Karrison, Theodore G, and others. 2016. "Versatile Tests for Comparing Survival Curves Based on Weighted Log-Rank Statistics." *Stata Journal* 16 (3). StataCorp LP: 678–90.

Pepe, Margaret Sullivan, and Thomas R Fleming. 1989. "Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data." *Biometrics*. JSTOR, 497–507.

———. 1991. "Weighted Kaplan-Meier Statistics: Large Sample and Optimality Considerations." *Journal of the Royal Statistical Society. Series B (Methodological)*. JSTOR, 341–52.

Péron, Julien, Marc Buyse, Brice Ozenne, Laurent Roche, and Pascal Roy. 2018. "An Extension of Generalized Pairwise Comparisons for Prioritized Outcomes in the Presence of Censoring." *Statistical Methods in Medical Research* 27 (4). SAGE Publications Sage UK: London, England: 1230–9.

Péron, Julien, Pascal Roy, Brice Ozenne, Laurent Roche, and Marc Buyse. 2016. "The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials." *JAMA Oncology* 2 (7). American Medical Association: 901–5.

Royston, Patrick, and Mahesh KB Parmar. 2011. "The Use of Restricted Mean Survival Time to Estimate the Treatment Effect in Randomized Clinical Trials When the Proportional Hazards Assumption Is in Doubt." *Statistics in Medicine* 30 (19). Wiley Online Library: 2409–21.

Uno, Hajime, B. Claggett, L. Tian, and et. al. 2014. "Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis." *Journal of Clinical Oncology*.