Robust Design and Analysis of Clinical Trials with Nonproportional Hazards: Methodology and Implementation with R

Satrajit Roychoudhury [1] and Keaven Anderson [2]

[1] Pfizer Inc. [2] Merck and Co., Inc.

ASA Biopharmaceutical Section Regulatory-Industry Statistics Workshop
September 22, 2020

Disclaimer

The views and opinions expressed herein are solely those of the presenters and are not necessarily those of Pfizer Inc. and Merck and Co. Inc. Any of these cannot and should not necessarily be construed to represent those of Pfizer Inc. and Merck and Co., Inc. or its affiliates.

Acknowledgement

- Pralay Mukhopadhyay, Astrazeneca Pharmaceuticals
- Jiabu Ye, Astrazeneca Pharmaceuticals
- Ji Lin, Sanofi
- Ray Lin, Roche
- Renee B Iacona, Astrazeneca Pharmaceuticals
- Tai-Tsang Chen, Glaxosmithkline
- Members of industry and FDA in the of cross-pharma non-proportional hazards working group
- Yilong Zhang, Merck and Co., Inc.
- Amin Shirazi, Iowa State University

Agenda

Welcome

1. Introduction: Motivation and Framework
2. Basics of Time to Event Analysis in Clinical Trial
3. Alternative Analysis Methods
4. Implementation using R- Part I

Break

5. Design Concepts for Time to Event Clinical Trial
6. Practical Designs in Presence of Non-proportional Hazard
7. Implementation using R- Part II
8. Designing TTE Trial with MaxCombo Test
9. Summary and Discussion

Learning Objective

- This course is for statistical researchers or students; personnel in the pharmaceutical industry, academic institutions, or regulatory agencies.
- Upon completing this course, participants will
  - Have better understanding of how to design and analyze time to event trial in presence of non-proportional hazard
  - Have familiarity with the R packages simtrial, gsDesign2, gsdmvn
  - Be able to intera9ct and communicate efficiently with relevant stakeholders

Course Material

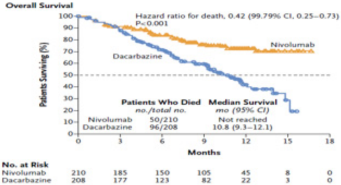
- Slides
  - lectures
  - practical
  - references
  - solutions
- R Packages: simtrial (K. M. Anderson (2020a)), gsDesign2 (K. M. Anderson (2020b)), gsdmvn K. M. Anderson (2020c)
- Detailed documentation: Design for the MaxCombo Test Under Non-Proportional Hazards

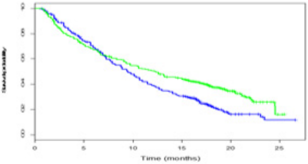
1. Introduction: Motivation and Framework

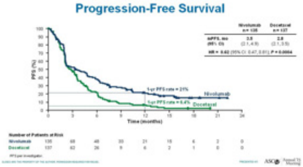
Non-Proportional Hazards (NPH): What Does It Mean?

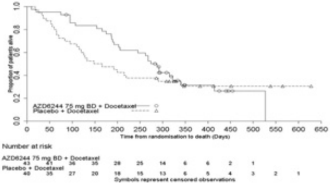
- Most popular methods in randomized clinical trial:
  - Kaplan-Meier (KM): describe chance of survival over time
  - log-rank test (LR test): detect difference in treatment effect (rejects "Null")
  - Cox regression: summarize the treatment effect
- Log-rank p-value, hazard ratio (HR), and naive median are the standard metrics of reporting
- Are they good summary measures when the treatment effect is not constant over time? : NPH problem
  - For example, recent immunotherapy development shows evidence of a delayed effect
- How to cope with NPH problem at design and analysis stages?

Recent Examples from Oncology Trials









Important Aspects of Design and Analysis with Potential NPH

- Analysis
  - Best method to use in presence of NPH
  - Analysis timing
  - Communication with broader audience and regulators
- Treatment effect quantifier
  - Underlying estimand
  - Relevance of HR
  - other options
- Trial design
  - Size and power of study
  - Interim analyses

2. Basics of Time to Event Analysis in Clinical Trial

Time to Event Analysis

- Time to Even (TTE) or Survival Analysis: methods to analyze time-to-event data
- Methods apply to the analysis of the magnitude or severity of a random event
- Terminology and emphases might differ in areas of application
    - TTE or Survival analysis: medicine, biology, public health (time to death)
    - Reliability analysis: engineering (time to a failure of some electronic component)
    - Duration analysis: economics (time looking for employment)
    - Severity analysis: finance (time to default)
    - Event history analysis: social sciences (time for doing some social and political task)

Goals of TTE Analysis

- Estimate TTE for a group of individuals: effect of treatment on risk of death
- Compare TTE between two or more groups: comparison between two treatments on risk of death **(Focus of this course)**
- Assess the relationship of covariates to TTE: relation between death and disease characteristics
- Relationship between multiple TTE variables: relationship between two endpoints (e.g., death and progression)

Notations and Terminology

A typical TTE data set contain patient level information for the following variables

- n_1 and n_0: number of subjects treated with experimental drug and control respectively; $n = n_1 + n_0$
- $\delta =$ Status for event of interest (e.g., death); $\sum_{j=1}^{n} \delta_i = D$
  - $\delta = 1 =>$ event observed; $X =$ time to event of interest
  - $\delta = 0 =>$ right censored; $U =$ time to right censoring or last known time to be event free
- $T = \min(X, U)$: Observed time
- $Z =$ covariates of interest: fixed or time-dependent

Right censoring is often seen in clinical studies as each patient is followed for pre-defined time period or cut-off time. Other censoring types (e.g., interval censoring) are also possible.

Survival and Hazard Function

- **Survival function**: $S(t) = P(T > t)$
  - Event free probability at time t
  - S is non-increasing with $S(0) = 1$ and $S(+\infty) = 0$

- **Hazard function**:

$$h(t) = \lim_{\Delta t -> 0} \frac{P(T \leq t + \Delta t | T > t)}{\Delta t}$$

  - *instantaneous risk* of the event happening at t given that it has not occurred before t
  - $h(t) > 0$ but it is not a probability

Cumulative Hazard Function

- **Cumulative Hazard function**:

$$H(t) = \int_0^\infty h(w)dw$$

  - Higher the value of $H(t)$, the greater the risk of failure by time t
  - Like the hazard function, the cumulative hazard function is not a probability
  - $H(t) \ = \ -\log S(t)$: referred to as *negative log survival*

Summary Measures

- **Milestone Survival at time** t_0: $S(t_0) = P(T > t_0)$
- **Mean survival time**: $\mu = E(T) = \int_0^\infty S(t)dt$
- **Restricted Mean survival time (RMST) to time L**: $\mu_L = E(\min(T, L)) = \int_0^L S(t)dt$
    - μ_L is more practical as μ may be large due to heavy tail
- **Percentile**: for $0 < q < 1$ the $100 \times q$-th percentile is

$$t_q = \inf\{t > 0 : S(t) \leq 1 - q)\}$$

    - Median survival time (50-th percentile): $m = \inf\{t > 0 : S(t) \leq 0.5)\}$

Nonparametric Methods

- Nonparametric estimation of $S(t)$
  - Kaplan-Meier (Product-limit) estimator
  - Breslow estimator
- No assumptions on the functional form of $S(t)$

Kaplan-Meier Estimator

- Idea is simple: based on discrete time and hazard
  - Depends on count only: *number at risk* and *number of events*
  - **Number of events** up to time t: $N(t) = I_{(T_i \leq t,\ \delta_i=1)}$
  - **Number at risk** at time t: $Y(t) = \sum_{i=1}^{n} I_{(T_i > t)}$ n= number of patients
- Kaplan-Meier (KM) Estimator= $\hat{S}_{KM}(t) = \prod_{u \leq t}(1 - \frac{\Delta N(u)}{Y(u)})$
  - Step function with jumps at event times
  - Inference about $S(t)$ is based on asymptotic normality of $\hat{S}_{KM}(t)$ (Klein et al. (2007))
  - All asymptotic properties are valid under independent or non-informative censoring

Estimation of Summary Measures

- **Estimated Mean survival time**: $\hat{\mu} = \int_0^\infty \hat{S}_{KM}(t)dt$

- **Estimated Restricted Mean survival time (RMST) to time L**: $\hat{\mu}_L = \int_0^L \hat{S}_{KM}(t)dt$

- **Estimated Percentile**: $\hat{t}_q = \inf\{t > 0 : \hat{S}_{KM}(t) \leq 1 - q)\}$
  - Estimated Median survival time: $\hat{m} = \inf\{t > 0 : \hat{S}_{KM}(t) \leq 0.5)\}$

Comparing Two Survival Curves at a Fixed Time Point

- Simplistic approach for comparing two survival curves
  - Appealing for it's simplistic clinical interpretation
- Two groups are compared at a pre-defined time point t_0 $(H_0 : S_0(t_0) = S_1(t_0))$using $\hat{S}_{KM}(t)$ and variance using Greenwood's formula (Klein et al. (2007))
- Substantial improvement of the properties of the test was obtained using proper transformations of the survival functions (e.g., $c \log \log$)
- Multiplicity adjustment is required if multiple time points are specified
- Depends heavily on the choice of t_0

Log-rank Test

- Popular test to test the null hypothesis of no difference in survival between two or more groups
- Adopted from stratified test for 2×2 contingency table
- Comparison based on the hazard functions not survival function
- The test can be written as

$$LR = \frac{\sum_{j=1}^{D}(O_j - E_j)}{\sqrt{\sum_{j=0}^{D} V_j}} = \frac{U}{se(U)}$$

O_j = Observed number of events and E_j= expected number of events at time t_j. V_j= variance of the observed number of events. Also U can be written as,

$$U = \int_0^T \left(dN_1(t) - Y_1(t)\frac{dN(t)}{Y(t)} \right)$$

Properties of Log-rank Test

- Rank based test
- LR is nonparametric in nature $=>$ no assumptions related to shape of survival function or treatment effect
- The power of LR depends on the number of observed events rather than the sample sizes
- Logrank test is most powerful for detecting the alternatives with constant treatment effect

$$H_1 : S_1(t) = S_0(t)^{exp\beta} \quad <=> h_1(t) = h_0(t)e^{\beta}$$

Regression Models for TTE

- Exploring association between covariates and survival time
- Main approaches include
  - **Proportional hazards model**: Most commonly used method in the analysis of TTE data

    $$h_1(t|z) = h_0(t)\exp(\beta z); \quad h_0(t)\text{:unspecified baseline hazard}$$

  - **Accelerated failure time (AFT) model**: Parametric model assumes accelerate or decelerate by covariate

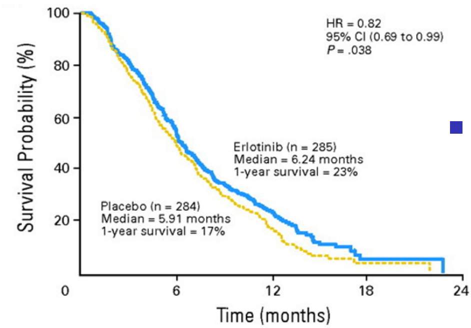
    $$\log T = z\beta + \sigma\epsilon \quad \epsilon\text{:fully specified distribution}$$

  - Other methods: Scale parameter model, frailty model, discreate time model etc.

Proportional Hazard (PH) Model

- Semiparametric: introduced by D.R. Cox 1972 (Cox (1972))
- Investigates the relationship of predictors and TTE through the hazard function
- Does not require assumption about underlying survival distribution
- Associated with log-rank test
- The effect of a covariate is described by hazard ratio (HR) $= e^{\beta}$: estimated using partial likelihood
    - Compares risk of event for the treatment group with control
    - Summary measures: point estimate of HR and 95% confidence interval (CI)
- **Proportional hazard or constancy of treatment effect is the key assumption**

Example Analysis of TTE (Moore et al. (2007))

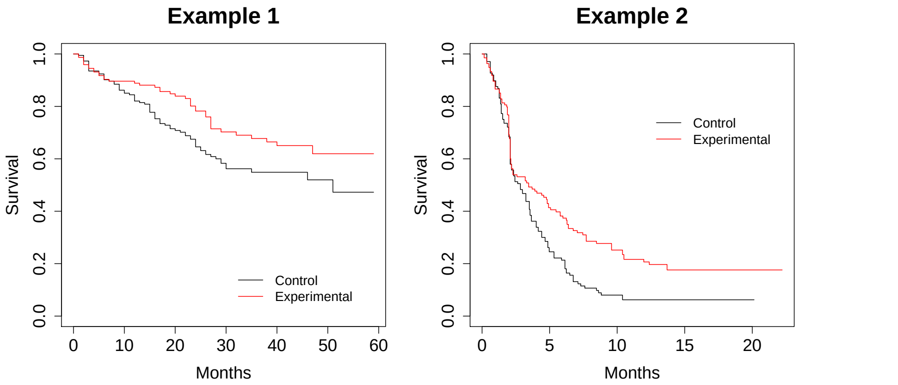


- Analysis of TTE includes: milestone survival rates, median, HR & 95% CI, and log-rank p-value

Assessing PH Assumptions

- Important to examine the proportional hazards assumption
    - Using statistical test and graphical diagnostics based on the scaled Schoenfeld residuals
    - Grambsch-Therneau (GT) test (Grambsch and Therneau (1994)): correlation between scaled Schoenfeld residuals and ranks of TTE
    - Recommend producing a graphical diagnostic
        - Schoenfeld residuals plot: non-random pattern against time confirms PH assumption
        - Cumulative hazard/log–log survival plots: plot of Nelson-Aalen estimates; PH assumption is reasonable if two plots are approximately parallel
    - Can be performed using the R package *survival*

Example


Example 1
Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
40
50
60
Months
Control
Experimental
Example 2
Survival
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
Months
Control
Experimental

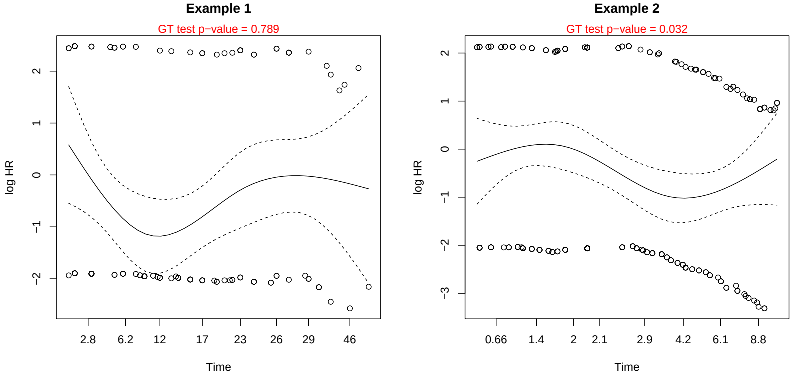
GT Test Using R

```r
library(survival)
library(survminer)
library(simtrial)

res.cox.1 <- coxph(Surv(stop,event)~rx, data = bladder)
test.ph.1 <- cox.zph(res.cox.1)


res.cox.2 <- coxph(Surv(month,evntd)~trt,
                   data = Ex2delayedEffect)
test.ph.2 <- cox.zph(res.cox.2)
```

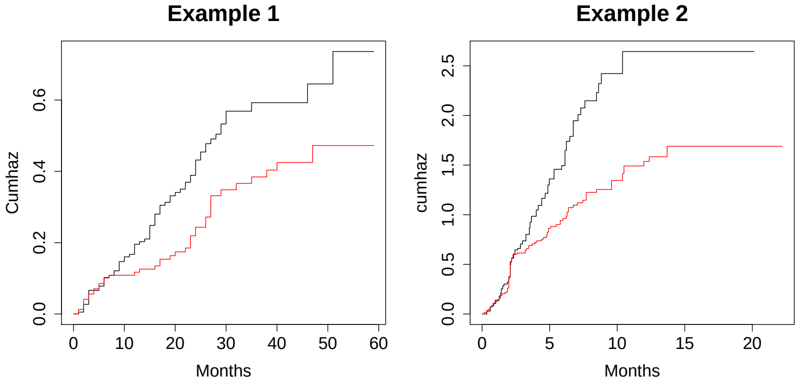
Schoenfeld Residuals Plot



Example 1

Example 2

Cumulative Hazard Plot vs Time

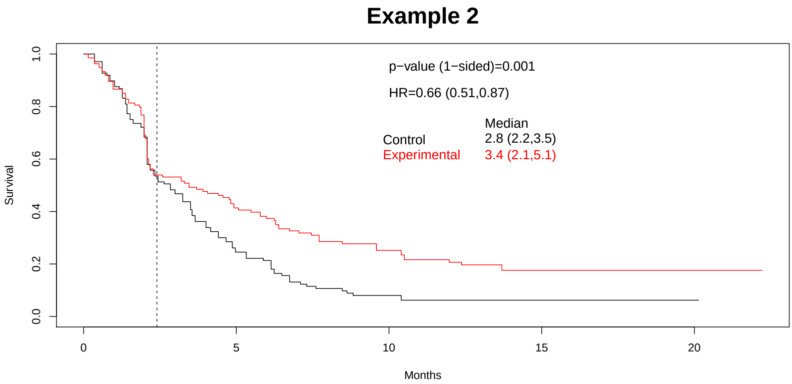


Example 1

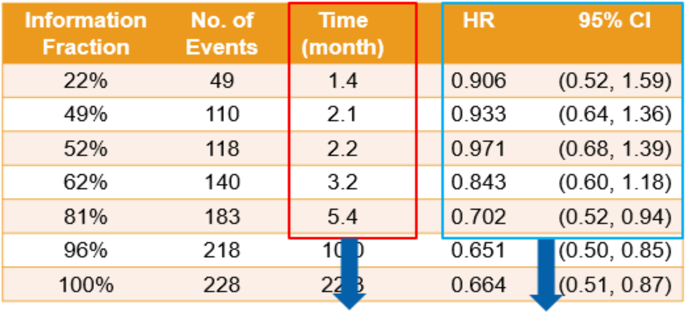
Example 2

3. Alternative Analysis Methods

Example 2 Revisited



Example 2: Treatment Effect Emerges Late in the Trial

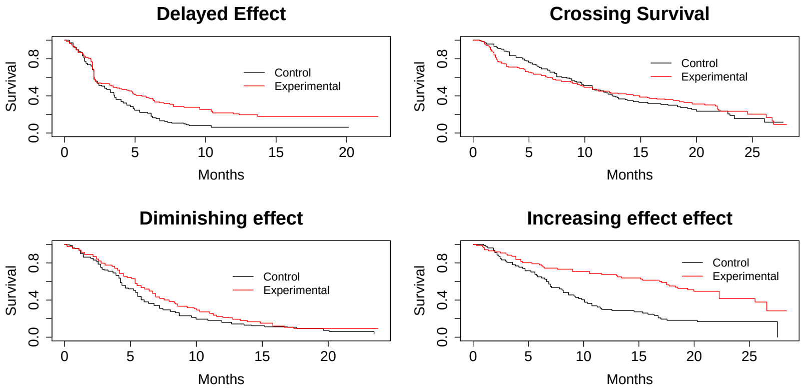


Information Fraction	No. of Events	Time (month)	HR	95% CI
22%	49	1.4	0.906	(0.52, 1.59)
49%	110	2.1	0.933	(0.64, 1.36)
52%	118	2.2	0.971	(0.68, 1.39)
62%	140	3.2	0.843	(0.60, 1.18)
81%	183	5.4	0.702	(0.52, 0.94)
96%	218	10.0	0.651	(0.50, 0.85)
100%	228	22.3	0.664	(0.51, 0.87)

Overall follow-up is low even with 80% events

Treatment effect emerges late in the trial

Different Types of Nonproportional Hazard (NPH)



Delayed Effect

Crossing Survival

Diminishing effect

Increasing effect effect

Key Challenges of Design and Analysis

- NPH has been discussed extensively in the survival analysis literature
    - Different methods for hypothesis testing and estimation are proposed
    - Methods are sensitive to the types of NPH
- $\sim$98% trials use log-rank test and Cox PH model for design primary analysis (source: NEJM 2000-2017)
    - Regulatory acceptable standard test and treatment effect summary
- **Main Challenges**: Uncertainty of NPH type at the design stage when PH assumption
    - The nature of treatment effect is unknown at the time of study design
    - A suitable design and analysis method must be handle multiple NPH types
    - Efficiently communicate the results to non-statisticians

Overview of Available Methods

- Focused on the methods generally used in drug development
- Methodologies can be broadly categorized as
  - Rank based
    - Weighted LR test, modestly weighted LR test, piecewise LR test
  - Kaplan-Meier based
    - Kaplan-Meier test (WKM), restricted mean survival time
  - Time dependent Cox regression (CoxTD)
  - Combination Test

Weighted Log-rank Test (WLR)

- LR test assumes that every point in time has the same relevance
- This assumption is questionable when treatment effect is not constant
- WLR attach a weight w_j with each points

$$WLR = \frac{\sum_{j=1}^{D} w_j(O_j - E_j)}{\sqrt{\sum_{j=1}^{D} w_j^2 V_j}} = \frac{U(w)}{se(U(w))}$$

$$U(w) = \int_0^T w(t)\left(dN_1(t) - Y_1(t)\frac{dN(t)}{Y(t)}\right)$$

  - LR test: $w_j = 1$
  - Wilcoxon (Gehan) test: $w_j = Y(t_j)$
  - Tarone-Ware test:$w_j = \sqrt{Y(t_j)}$
  - Several others . . .

Flemming-Harrington WLR

- Fleming and Harrington proposed a class of weighted log-rank test (FH) based on the $G^{\rho,\gamma}$ family
  - Covers wide variety of treatment effect scenarios with appropriate choice of ρ and γ
- The weights are provided using the formula

$$w(t) = (\hat{S}(t-))^{\rho}(1 - \hat{S}(t-))^{\gamma} \ (\rho \geq 0, \gamma \geq 0)$$

  $\hat{S}(t-)$ is the estimated survival function immediately prior to time t
- Values of ρ and γ can handle different treatment effect types:
  - $\rho > 0$, $\gamma = 0$: early difference
  - $\rho = 0$, $\gamma > 0$: late difference
  - $\rho > 0$, $\gamma > 0$: mid difference
  - $\rho = 0$, $\gamma = 0$: log-rank test ("equal weighting")

Weighted Hazard Ratio (WHR)

- Cox hazard ratio can be interpreted as an average over the observed event times (Grambsch and Therneau (1994))
- A dual estimate of treatment effect quantifier for the WLR test is the weighted hazard ratio (WHR)
  - Time averaged hazard ratio using the weights are same as the associated WLR test
- Estimated using weighted partial likelihood of Cox model (Schemper, Wakounig, and Heinze (2009))

$$\sum_{j=1}^{D} w(t_j) \frac{\partial l_j}{\partial \beta} = 0$$

  l_j: the log partial likelihood of Cox model and $w(t_j)$ weight for WLR
- Confidence interval can be calculated using asymptotic properties

Challenges with WLR an WHR

- The choice of ρ and γ requires knowledge of the shape of survival curves and plays an important role to the performance of WLR test and WHR
  - Mis-specification of the weight function may result in loss of power
- WHR often lacks intuitive interpretation
  - The associated estimad is complex
  - Often lacks causal interpretation
  - Hard to explain to non-statisticans

Modestly Weighted LR Test (mWLRT)

- Another version of the WLR test introduced by Magirr and Burman (2019)
- Based on the score test representation of WLR
- Maggir and Burman 2019 proposed a test with nonincreasing scores but increasing weights
  - The scores set to 1 for all $t \leq t^*$ (clinically meaningful timepoint, e.g., 12 months) and nonincreasing thereafter (similar to LR)
  - The is equivalent to setting $w(t) = 1/\max(\hat{S}(t), \hat{S}(t^*))$; $\hat{S}(t)$ denotes the Kaplan-Meier estimate of the pooled data at time t
- Simulation studies showed mWLRT
  - Protects the type-I error under strong null $(H_0^{Strong} : S_1(t) \leq S_0(t))$
  - Higher statistical power than LR under delayed treatment effect scenario
  - Comparable power with LR under PH scenario
- The corresponding WHR will be difficult to interpret for practical purposes

Piecewise LR test (pWLRT)

- Xu et al. (2017), Xu et al. (2018) propose pWLRT for two intervals;

$$pWLRT = \frac{\sum_{j \in D_1} w_1(O_j - E_j) + \sum_{l \in D_2} w_2(O_l - E_l)}{\sqrt{\sum_{j \in D_1} w_1^2 V_j + \sum_{l \in D_2} w_2^2 V_l}}$$

  D_1 and D_2: indices of patients who had event before and after t^*

- Power/Type-I error can be calculated analytically or using simulation
- Most powerful under delayed treatment effect when $w_1 = 0$: "ignores early events"
- Depends heavily on the specification of t^*
- A piecewise HR captures the time dependence nature of treatment effect

Weighted Kaplan-Meier Test (WKM)

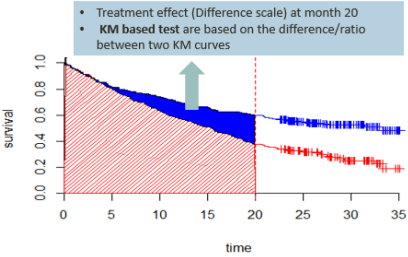
- Class of distance tests introduced by Pepe and Fleming (1989)
- Based on the weighted KM statistic of two groups, and integrating over the restricted range after a specified cut-off

$$WKM = \int_0^\tau \sqrt{\frac{n_1 n_0}{n}} \hat{w}(t)[\hat{S}_{1KM}(t) - \hat{S}_{0KM}(t)]dt$$

  $\hat{w}(t)$ is geometric average of the two censoring survivor function estimators and τ is the largest follow-up time
- Asymptotic properties are derived by Pepe and Fleming (1989)
- Corresponding treatment effect does not have intuitive interpretation except $\hat{w}(t) = 1$
- Dependent of the choice of τ

Restricted Mean Survival Time (RMST)

- Area under the KM plot prior to specific time-point τ ($\hat{w}(t) = 1$)
- Treatment effect estimator: difference or ratio of RMST: can be easily interpreted as “life expectancy” (Royston and Parmar (2011), Uno et al. (2014))
- Performance of RMST depends on censoring pattern and choice of τ
  - Data-dependent: unknown at the design state



- Treatment effect (Difference scale) at month 20
- **KM based test** are based on the difference/ratio between two KM curves

Cox Regression with Time Dependent Coefficient (CoxTD)

- A natural extension of Cox regression model for NPH setting is including a time varying coefficient for treatment (Putter et al. (2005))
$$h(t) = h_0(t)\exp(Z\beta_F + Zf(t)\beta_T)$$
- $\log(t+1)$ as a "reasonable" choice for $f(t)$ to diminish the influence of very early events
- Likelihood ratio test for $H_0 : S_1(t) = S_0(t)$
- Simulation shows that CoxTD model does not perform well in terms of power under delayed treatment effect (Callegaro and Spiessens (2017))
- Reporting the HR as a continuous function is hard to interpret by nonstatisticians

Combination Test

- Handle a broad class of alternative hypothesis: Lee (2007), Karrison and others (2016), Breslow, Edler, and Berger (1984)
- Considers multiple test statistics: choose best test statistics based on data
  - Breslow, Edler, and Berger (1984): combination of LR test and test of acceleration
  - Logan, Klein, and Zhang (2008): combination of LR test and milestone survial
  - Lee (2007): Average and maximum of LR test ($FH(0,0)$) and $FH(0,1)$
- Requires appropriate multiplicity control due to the correlation of test statistics
- Often provides robust power under wide class of alternative hypotheses
- Communication of treatment effect is often difficult due to complex nature

Other Methods

- Net benefit} or the net chance of a longer event-free (Buyse (2010), Perón et al. (2016))
    - Generalized pairwise difference: probability that a random patient in the treatment group is event-free by at least a pre-specified difference as compared to a random patient in the control group minus the probability of the reverse situation
    - Under PH, Net benefit = [1-HR]/[1+HR]
- AFT model
- Change-point model

Choice of Primary Analysis

- Regarding primary analysis ICH E9 states

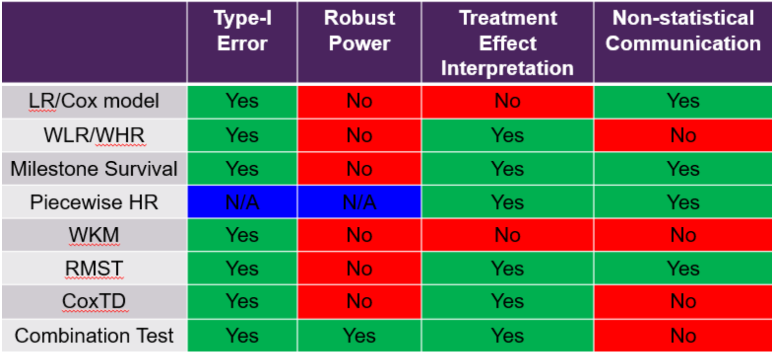
For each clinical trial contributing to a marketing application, all important details of its design and conduct and the principal features of its proposed statistical analysis should be clearly specified in a protocol written before the trial begins. The extent to which the procedures in the protocol are followed and the primary analysis is planned a priori will contribute to the degree of confidence in the final results and conclusions of the trial.

- Specifying primary analysis when NPH is expected: need robust statistical method to handle
  - Possibility of different types of NPH
  - Possibility of different specifications (e.g. lag time for treatment effect)

A Qualitative Evaluation

- A primary analysis involves both testing and estimation of treatment effect
- We perform a qualitative of avalable methods based on 4 important metrics
    - **Type-I error**: Controlling type I error at a specific level of significance (e.g., 2.5%) under the null hypothesis H_0: $S_1(t) = S_0(t)$ for all t.
    - **Robust power**: Showing resilience in terms of statistical power when the PH assumption is violated. Often a statistical test suffers a power loss when the nature of the underlying treatment effect is not anticipated
    - **Treatment effect Interpretation**: Interpretable treatment effect summary under various types of PH and NPH
    - **Non-statistical Communication**: Easy to understand by non-statisticians

Qualitative Review Under NPH



	Type-I Error	Robust Power	Treatment Effect Interpretation	Non-statistical Communication
LR/Cox model	Yes	No	No	Yes
WLR/WHR	Yes	No	Yes	No
Milestone Survival	Yes	No	Yes	Yes
Piecewise HR	N/A	N/A	Yes	Yes
WKM	Yes	No	No	No
RMST	Yes	No	Yes	Yes
CoxTD	Yes	No	Yes	No
Combination Test	Yes	Yes	Yes	No

Potential Candidates for Confirmatory Trial

- Under NPH, no single efficacy measure is sufficient
- Milestone survival, RMST, CoxTD, and combination tests are potential candidates
- However, WKM, CoxTD, milestone survival and RMST fail to show robust power under a wide class of alternatives (Lin et al. (2020), Callegaro and Spiessens (2017))
- An improvement over the available tests and provides robust power
- If NPH is not expected, we recommend the use of traditional LR test and HR for the primary analysis

Robust MaxCombo Test

- Proposed by Cross-Industry NPH working group (Roychoudhury et al. (2020), Lin et al. (2020))
- Motivated from the work from Yang and Prentice (2010) and Lee (2007)
- Based on multiple FH-WLR test statistics and chooses the best one adaptively depending on the underlying data
- We consider two possible combination tests
  - **MaxCombo**: FH(0,0),FH(0,1), FH(1,1), FH(1,0)
  - **Modified MaxCombo**: FH(0,0),FH(0,0.5), FH(0.5,0.5), FH(0.5,0)
  - Other candidate: FH(0,0),FH(0,0.5), FH(0.5,0.5)
- Able to handle PH, delayed effectect, crossing survival, early-separation, and mixture of more than one NPH type scenarios as alternative

Null Distribution of MaxCombo Test

- The proposed combination test

$$Z_{max} = max\{FH(\rho_i, \gamma_j):\ (\rho_i, \gamma_j) = (0,0), (0,1), (1,0), (1,1)\}$$

- The type I error and power calculation require the joint distribution of four FH-WLR test statistics
- Karrison and others (2016) proved that the joint distribution is asymptotically normal

$$(FH(0,0), FH(0,1), FH(1,0), FH(1,1)) \sim N_4(\mathbf{0},\ \mathbf{\Gamma})\quad \text{under} H_0$$

Calculation of p-Value

- With correlation matrix $\Gamma = ((\eta_{ij}))$ is of the following form;

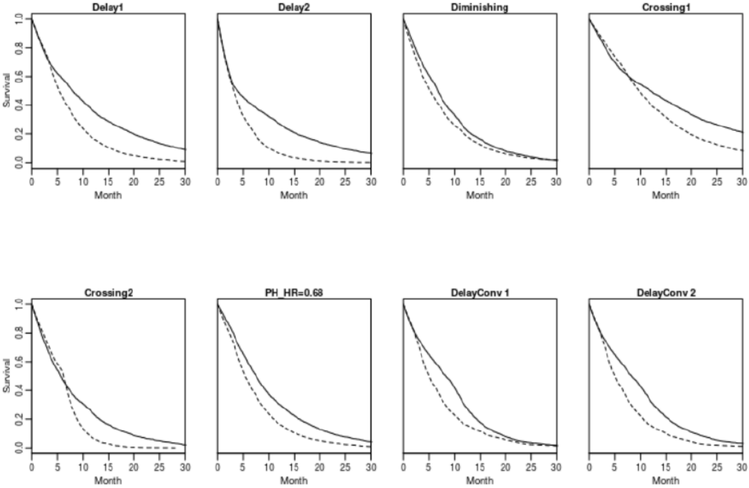
$$\eta_{ij} = \frac{\text{Cov}(FH(\rho_i, \gamma_i), FH(\rho_j, \gamma_j)}{\sqrt{V(FH(\rho_i, \gamma_i))V(FH(\rho_j, \gamma_j}))} \\ = \frac{V(FH(\frac{\rho_i+\rho_j}{2}, \frac{\gamma_i+\gamma_j}{2}))}{\sqrt{V(FH(\rho_i, \gamma_i))V(FH(\rho_j, \gamma_j))}} \quad \text{for } i \neq j$$

- One-sided p-value calculation uses multivariate normal calculation:

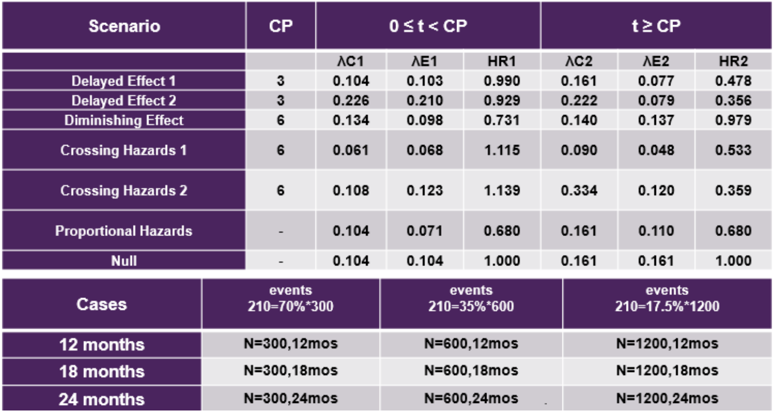
$$p - value = P(Z_{max} > z_{max} | H_0) \\ = 1 - \int_{-\infty}^{z_{max}} \int_{-\infty}^{z_{max}} \int_{-\infty}^{z_{max}} \int_{-\infty}^{z_{max}} \phi_4(\boldsymbol{\omega}, \mathbf{0}, \boldsymbol{\Gamma}) d\boldsymbol{\omega}$$

- Calculation can be done using efficient integration routine in R and SAS (Genz (1992))

Simulation Study (Lin et al. (2020))



Wide Number of NPH Scenarios Considered

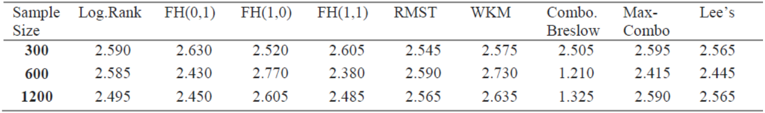


Scenario	CP	0 ≤ t < CP			t ≥ CP		
		λC1	λE1	HR1	λC2	λE2	HR2
Delayed Effect 1	3	0.104	0.103	0.990	0.161	0.077	0.478
Delayed Effect 2	3	0.226	0.210	0.929	0.222	0.079	0.356
Diminishing Effect	6	0.134	0.098	0.731	0.140	0.137	0.979
Crossing Hazards 1	6	0.061	0.068	1.115	0.090	0.048	0.533
Crossing Hazards 2	6	0.108	0.123	1.139	0.334	0.120	0.359
Proportional Hazards	-	0.104	0.071	0.680	0.161	0.110	0.680
Null	-	0.104	0.104	1.000	0.161	0.161	1.000

Cases	events 210=70%*300	events 210=35%*600	events 210=17.5%*1200
12 months	N=300,12mos	N=600,12mos	N=1200,12mos
18 months	N=300,18mos	N=600,18mos	N=1200,18mos
24 months	N=300,24mos	N=600,24mos	N=1200,24mos

Simulation Results: Type-I Error

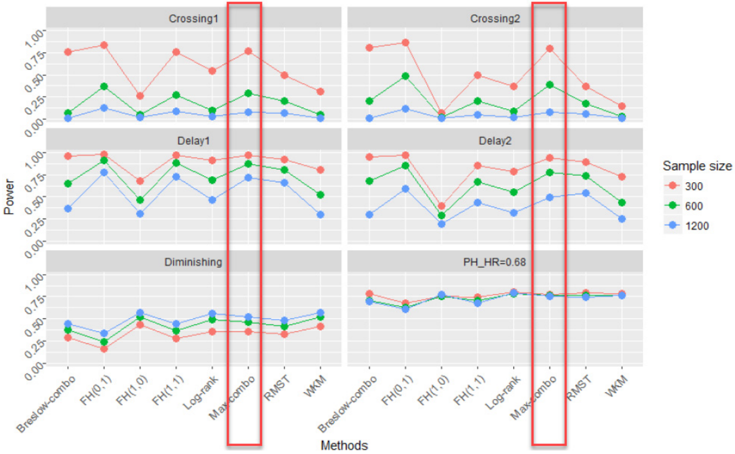
- Two additional scenarios with delayed effect with converging tails
- 20,000 trial datasets are simulated for each scenario
- Type-I error is well protected with MaxCombo test with null $H_0 : S_1(t) = S_0(t)$ for all scenarios



Sample Size	Log.Rank	FH(0,1)	FH(1,0)	FH(1,1)	RMST	WKM	Combo. Breslow	Max-Combo	Lee's
300	2.590	2.630	2.520	2.605	2.545	2.575	2.505	2.595	2.565
600	2.585	2.430	2.770	2.380	2.590	2.730	1.210	2.415	2.445
1200	2.495	2.450	2.605	2.485	2.565	2.635	1.325	2.590	2.565

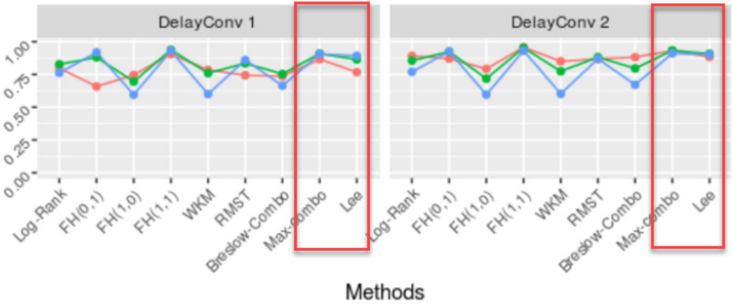
Simulation Results: Power

- Robust power across different NPH scenarios
- 3-4% power loss under PH scenario



Advantage Over Existing Combination Test

- MaxCombo test has improved power over Lee test under delayed effect with converging tails scenarios



Further Criticism of MaxCombo Test

- Additional simulations are performed to address further criticism about $FH(0,1)$ and MaxCombo test (Freidlin and Korn (2019), Magirr and Burman (2019))

- There are some concerns regarding the performace of the MaxCombo test under the *strong null* and *severe late crossing* scenarios

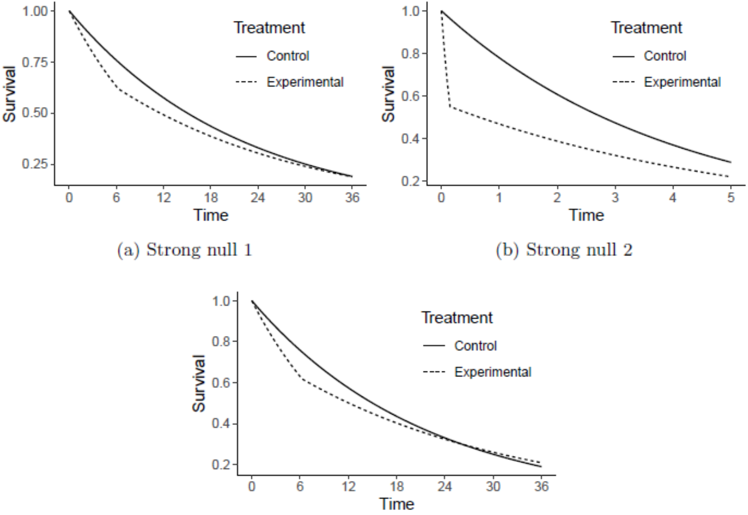
  - Possibility of high probability of rejecting null hypothesis when the experimental drug is actually harmful

- We have considered the following three scenarios

  - *Strong Null 1* : Magirr and Burman (2019)
  - *Strong Null 2* : Freidlin and Korn (2019)
  - *severe late crossing*: the treatment group shows a late and marginal survival benefit over the control group which makes the overall treatment effect clinically questionable

- Should not be mixed with type-I error assessment

MaxCombo Under Extreme Scenarios (Roychoudhury et al. (2020))



(a) Strong null 1

(b) Strong null 2

(c) Severe late Crossing

Results

- The final cut-off date for each simulation is the calendar time of 5 years
  - All patients alive at that point are censored at the cut-off date
- Probability of rejecting null hypothesis is **low with MaxCombo test** under strong null 1 and severe crossing scenarios
  - Recruitment uniformly over 12 months: **2.1%** (strong null 1); **5.0%** (severe crossing)
  - Recruitment uniformly over 6 months: **2.3%** (strong null 1); **5.8%** (severe crossing)
- Probability of rejecting null hypothesis is high for scenario 2 (48.9 %)
  - Can be handled using alternative weighting scheme: **Modified MaxCombo test reduced this probability to 1.8%**
  - **Modified MaxCombo test** also handles the severe crossing scenario well (2.6%)
  - Such scenarios are unrealistic in real-life: will stopped early by a data monitoring committee (DMC) due to the safety concerns
- **MaxCombo and Modified MaxCombo tests showed better power that LR and mWLRT** under delayed effect and crossing survival

Estimation of Treatment Effect

- The dual WHR of MaxCombo test is calculated based on the best weight chosen
- Estimated using weighted Cox regression
- 95% CI calculation requires the joint distribution of FH-WLR test statistics
- $100 \times (1-\alpha)\%$ simultaneous confidence interval corresponding for WHR related to MaxCombo can be calculated as

$$\hat{HR}^{MaxCombo} \pm C^{*} \times SE(\hat{HR}^{MaxCombo})$$

  C^{*} is calculated using the asymptotic multivariate normal distribution of WHR (Karrison and others (2016)})

- However, the WHR has limited interpretation to non-statisticians

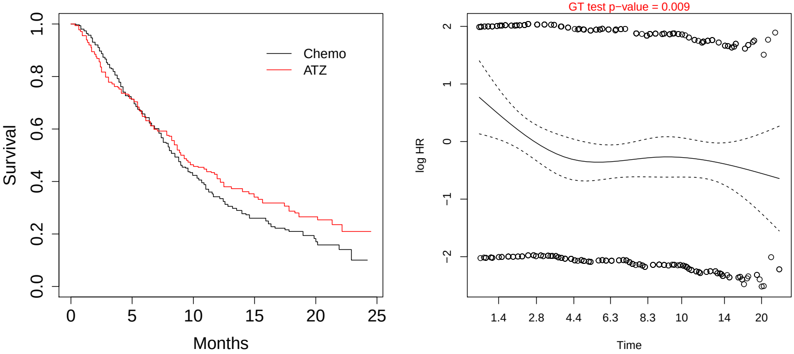
Primary Analysis for Confirmatory Trials

- Under NPH, no single efficacy measure is sufficient
- A p-value from any single statistical test or a single summary statistic fails to capture treatment benefit
- A robust testing procedure like MaxCombo or modified MaxCombo test is required to handle uncertainties associated with NPH type
- Additional pre-specified measures beyond HR and median needed to describe benefit over entire follow-up period; e.g., milestone survival, RMST
- Important to ensure adequate follow-up to evaluate time-dependent treatment effect

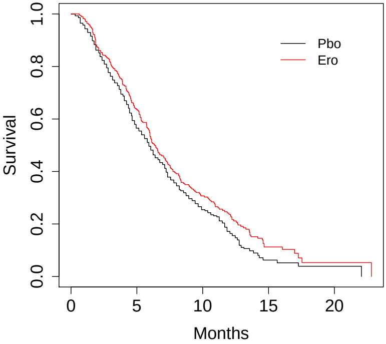
Specification of Primary Analysis in Protocol

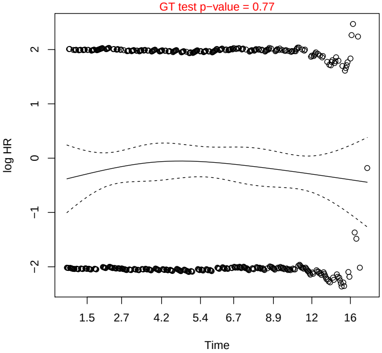
- A stepwise approach for primary analysis in trials where NPH is expected
  - **Step 1**: Perform a statistical test to reject "Null" hypothesis (no treatment effect) using MaxCombo or modified MaxCombo test
  - **Step 2**: Evaluate PH assumption using standard methods
  - **Step 3**: Select treatment effect summary based on step 2 findings
    - if PH is reasonable: use traditional measures like HR and median
    - if PH is not reasonable: also provide additional measures such as milestone survival rate, RMST, and piecewise HR at pre-specified time points
- This approach provides a complete summary of any treatment effect
- Appropriately pre-specification is possible to meet ICH E9

Example 1: Overall Survival IM211 Trial IC1/2/3 Cohort (Digitized)

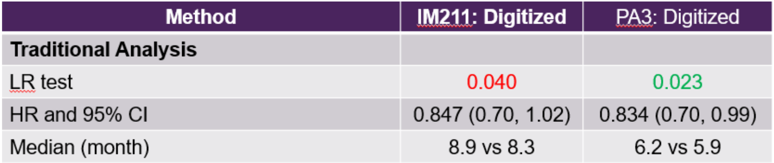


Example 2: Overall Survival PA3 Trial (Digitized)



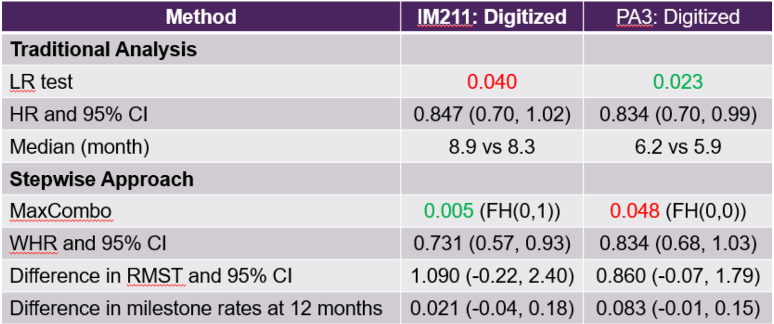


Use of Stepwise Approach



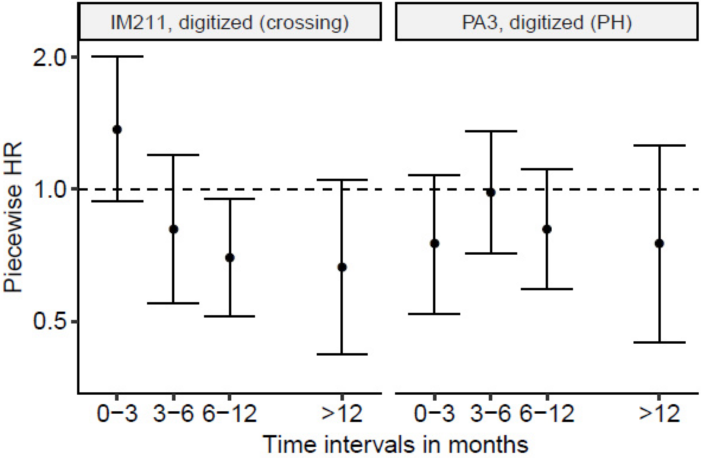
Method	IM211: Digitized	PA3: Digitized
Traditional Analysis		
LR test	0.040	0.023
HR and 95% CI	0.847 (0.70, 1.02)	0.834 (0.70, 0.99)
Median (month)	8.9 vs 8.3	6.2 vs 5.9

Use of Stepwise Approach



Method	IM211: Digitized	PA3: Digitized
Traditional Analysis		
LR test	0.040	0.023
HR and 95% CI	0.847 (0.70, 1.02)	0.834 (0.70, 0.99)
Median (month)	8.9 vs 8.3	6.2 vs 5.9
Stepwise Approach		
MaxCombo	0.005 (FH(0,1))	0.048 (FH(0,0))
WHR and 95% CI	0.731 (0.57, 0.93)	0.834 (0.68, 1.03)
Difference in RMST and 95% CI	1.090 (-0.22, 2.40)	0.860 (-0.07, 1.79)
Difference in milestone rates at 12 months	0.021 (-0.04, 0.18)	0.083 (-0.01, 0.15)

Piecewise HR



4. Implementation using R- Part I

Packages used

- survival package
    - Kaplan-Meier survival estimates
    - Cox model
    - logrank testing
    - Limited use for weighted logrank
- simtrial
    - Example datasets
    - Counting process data model
    - Weighted logrank tests
    - Combination tests (MaxCombo)
- dplyr
    - tidy data manipulation

Installing Packages

```r
devtools::install_github("keaven/simtrial")
devtools::install_github("keaven/gsDesign2")
devtools::install_github("keaven/gsdmvn")
```

Delayed effect dataset - Introduction

```
head(Ex2delayedEffect, n= 3) %>% kable(digits=3)
```

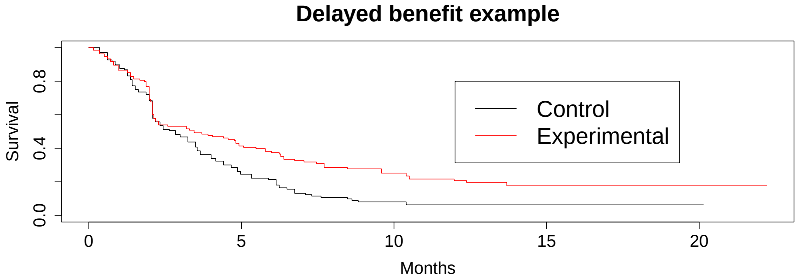
id	month	evntd	trt
1	0.152	1	1
2	0.152	1	1
3	0.355	1	1

```
with(Ex2delayedEffect, table(evntd, trt))
```

```
##       trt
## evntd   0   1
##     0  14  30
##     1 123 105
```

Delayed effect dataset - Plotting

```r
plot(survfit(Surv(month,evntd)~trt, data = Ex2delayedEffect),
     col=1:2, ylab = "Survival", xlab = "Months",
     main = "Delayed benefit example",
     cex.main = 2, cex.lab = 1.5, cex.axis = 1.5)
legend(x = 12, y = .8, legend = c("Control", "Experimental"),
       col=1:2, lty=1, cex = 2)
```



Delayed effect dataset - Cox model

- `exp(coef)` is hazard ratio (HR) for this binary model
- p-value is 2-sided (Chi-square version of logrank)

```r
fit <- coxph(Surv(month,evntd)~trt,
             data = Ex2delayedEffect)
fit
```

```
## Call:
## coxph(formula = Surv(month, evntd) ~ trt, data = Ex2delayedEf
##
##        coef exp(coef) se(coef)      z      p
## trt -0.4093    0.6641   0.1354 -3.024 0.0025
##
## Likelihood ratio test=9.19  on 1 df, p=0.002435
## n= 272, number of events= 228
```

Delayed effect dataset - Test for NPH

- In this case, Grambsch-Therneau test {*shows? suggests?*} *a difference*
- *Generally, this (any) test for NPH is underpowered*

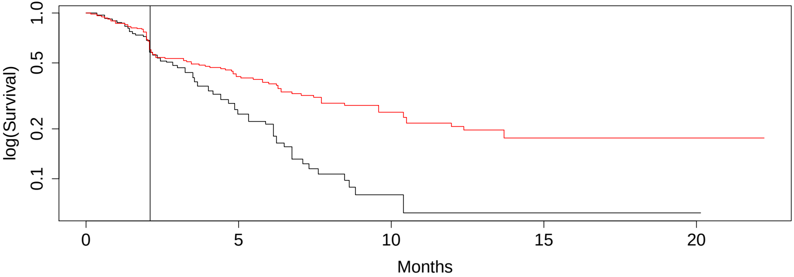
```
cox.zph(fit)
```

```
##       rho chisq      p
## trt -0.14  4.61 0.0318
```

Delayed effect dataset - exponential failure?

- Plot survival on log scale
- Slope is hazard rate; constant? piecewise linear?

```
plot(survfit(Surv(month,evntd)~trt, data = Ex2delayedEffect),
  col=1:2, ylab = "log(Survival)", xlab = "Months", log = "y",
  cex.lab = 1.5, cex.axis = 1.5)
abline(v=2.1)
```



Piecewise Cox model - delayed effect example

First 2.1 months

```
coxph(Surv(month,evntd)~trt,
  data = Ex2delayedEffect %>%
  mutate(le21 = (month <= 2.1) * 1, evntd = le21 * evntd))
```

```
## Call:
## coxph(formula = Surv(month, evntd) ~ trt, data = Ex2delayedEf
##     mutate(le21 = (month <= 2.1) * 1, evntd = le21 * evntd))
##
##         coef exp(coef) se(coef)      z     p
## trt -0.06921   0.93313  0.19084 -0.363 0.717
##
## Likelihood ratio test=0.13  on 1 df, p=0.7168
## n= 272, number of events= 110
```

Piecewise Cox model - delayed effect example

After 2.1 months

```
coxph(Surv(month,evntd)~trt,
      data = Ex2delayedEffect %>%
        filter(month > 2.1))
```

```
## Call:
## coxph(formula = Surv(month, evntd) ~ trt, data = Ex2delayedEf
##     filter(month > 2.1))
##
##        coef exp(coef) se(coef)      z        p
## trt -0.7361    0.4790   0.1895 -3.884 0.000103
##
## Likelihood ratio test=15.2  on 1 df, p=9.66e-05
## n= 157, number of events= 118
```

Failure rates by period - delayed effect example

Controls; constant rate over time?

```r
with(Ex2delayedEffect %>% filter(trt == 0),
     pwexpfit(Surv(month, evntd),intervals=2.1)) %>%
  kable(digits=3) %>% kable_styling()
```

intervals	TTOT	events	rate	m2ll
2.1	246.938	57	0.231	281.134
Inf	301.523	66	0.219	332.533

Experimental: no effect early, HR $\sim$ 0.5 late

```r
with(Ex2delayedEffect %>% filter(trt == 1),
     pwexpfit(Surv(month, evntd),intervals=2.1)) %>%
  kable(digits=3) %>% kable_styling()
```

intervals	TTOT	events	rate	m2ll
2.1	245.690	53	0.216	268.580
Inf	594.488	52	0.087	357.392

Testing with logrank

Approximately the same as Wald test from earlier Cox model

```
survdiff(Surv(month, evntd) ~ trt, data = Ex2delayedEffect)
```

```
## Call:
## survdiff(formula = Surv(month, evntd) ~ trt, data = Ex2delaye
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## trt=0 137      123      101      4.82      9.25
## trt=1 135      105      127      3.83      9.25
##
##  Chisq= 9.3  on 1 degrees of freedom, p= 0.002
```

If stratifying by, say, sex:

```
survdiff(Surv(month, evntd) ~ trt + strata(sex),
         data = Ex2delayedEffect)
```

Testing with simtrial

- Targets logrank, weighted logrank, MaxCombo tests
- Requires fixed variable names in survival data (ugh!)
    - No model statement
- Sets up counting process interim dataset
- Pre-set or user-defined weighting for weighted logrank

Changing variable names

```r
ex2 <- Ex2delayedEffect %>%
       transmute(Stratum = "All",
                 Treatment = trt,
                 tte = month,
                 event = evntd)
head(ex2, n=5) %>% kable(digits=3) %>% kable_styling()
```

Stratum	Treatment	tte	event
All	1	0.152	1
All	1	0.152	1
All	1	0.355	1
All	1	0.355	1
All	1	0.355	1

Translate to counting process dataset

- sorted by tte (time-to-event)
- only records for times with events

```
# txval is indicator of experimental treatment
ex2counting <- ex2 %>% tensurv(txval = 1)
head(ex2counting, n = 5) %>% kable(digits=2) %>%
  kable_styling(font_size = 8)
```

Stratum	events	txevents	tte	atrisk	txatrisk	S	OminusE	Var
All	2	2	0.15	272	135	1.00	1.01	0.50
All	7	3	0.36	270	133	0.99	-0.45	1.71
All	2	2	0.51	263	130	0.97	1.01	0.50
All	8	2	0.61	260	127	0.96	-1.91	1.94
All	2	2	0.71	252	125	0.93	1.01	0.50

What are the counting process variables?

- `Stratum` - stratum (discrete values)
- `tte` - time at which event(s) occurred
- `events` - number of events at time `tte`
- `txevents` - number of events in experimental group
- `atrisk` - number at risk just before time `tte`
- `txatrisk` - number at risk in experimental group just before `tte`
- `S` - Kaplan-Meier survival (left-continuous!) at time `tte`; overall population
- `OminusE` - Observed events minus expected for experimental if no treatment effect
- `Var` - variance of `OminusE` (hypergeometric)

Defining a weight

- Weight first 2.1 months is 0
- This is a one-sided test

```
ex2counting <- ex2counting %>% mutate(w= (tte > 2.1) * 1)
ex2counting %>% ungroup() %>%
  summarize(numerator = sum(OminusE * w),
            denominator = sqrt(sum(w^2 * Var)),
            Z = numerator / denominator,
            p = pnorm(Z))  %>% kable() %>% kable_styling()
```

numerator	denominator	Z	p
-20.24563	5.137288	-3.940918	4.06e-05

Using tenFH() for logrank

- Z-test
- rho=0, gamma=0 indicate logrank
- p-value is 2-sided, as before

```
ex2counting %>% tenFH(rg = tibble(rho=0, gamma=0)) %>%
  mutate(pnorm(Z) * 2)  %>% kable(digits=3) %>% kable_styling()
```

rho	gamma	Z	pnorm(Z) * 2
0	0	-3.042	0.002

Some Fleming-Harrington tests

- One-sided weights
- (rho = 0, gamma = 0.5) and (rho = 0.5, gamma = 0.5) often good options!
  - not too much down-weighting

```
rg <- tibble(rho =   c(0,  0, 0, .5, 1),
             gamma = c(0, .5, 1, .5, 1))
ex2counting %>% tenFH(rg = rg) %>%
  mutate(p = 2 * pnorm(Z),
   test=c("logrank", "down-weight early", "down-weight early",
          "up-weight middle", "up-weight middle")) %>%
  kable(digits=c(1,1,3,5,0)) %>% kable_styling()
```

rho	gamma	Z	p	test
0.0	0.0	-3.042	0.00235	logrank
0.0	0.5	-3.671	0.00024	down-weight early
0.0	1.0	-3.792	0.00015	down-weight early
0.5	0.5	-3.408	0.00065	up-weight middle
1.0	1.0	-3.488	0.00049	up-weight middle

MaxCombo test

```r
# use logrank, (rho = 0, gamma= .5), (rho = .5, gamma = .5)
rg = tibble(rho = c(0, 0, .5), gamma = c(0, .5, .5))
Z <- ex2counting %>% tenFHcorr(rg = rg)
Z %>% kable(digits=3)
```

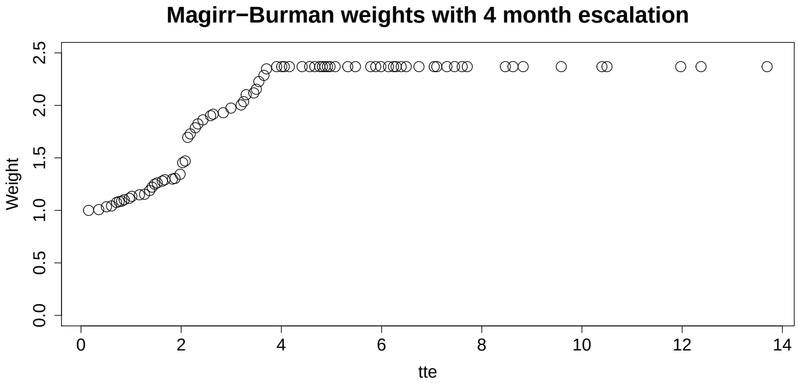
rho	gamma	Z	V1	V2	V3
0.0	0.0	-3.042	1.000	0.933	0.967
0.0	0.5	-3.671	0.933	1.000	0.972
0.5	0.5	-3.408	0.967	0.972	1.000

```r
# NOTE: one-sided
pMaxCombo(Z)
```

```
## [1] 0.0001211726
```

Magirr-Burman test

```
# Down-weight for 4 months
MBcounting <- ex2counting %>% wMB(delay = 4)
```



Magirr-Burman Modestly Weighted logrank

- Similar to logrank in this case
- Can be a nice alternative to logrank, Fleming-Harrington or MaxCombo
- Control of Type I error under strong null hypothesis

```
MBcounting %>% summarize(S = sum(OminusE*wMB),
                         V = sum(Var*wMB^2),
                         Z = S / sqrt(V),
                         p = 2 * pnorm(Z)) %>%
  kable(digits=c(1,2,2,6))
```

S	V	Z	p
-48.3	170.76	-3.69	0.00022

5. Design Concepts for Time to Event Clinical Trial

Basics of Design with TTE

- Event driven: timing of the analysis depends on targeted number of events
- Sample size is traditionally calculated using LR test (Schoenfeld (1981))
- Required number of events D for is calculated as:

$$D = \frac{(r+1)^2}{r} \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\theta^2}$$

r: randomization ratio; α: level of significance; $1-\beta$: required power; and θ: log of alternative hypothesis

- Calculate number of patients and follow-up time needed to observe D events

Example of Traditional Fixed Design

- Design set up: Treatment A vs SOC
  - Progression Free Survival (PFS) as primary endpoint
  - Median SOC: 5 months
  - Alternative HR (Treatment A vs SOC): 0.67
  - Enrollment period: 37.1 months
  - Type I error 2.5%, power 90%
- Requires:
  - 300 (150 per arm) patients
  - 263 events to target : analysis timing $\sim$ 44 months
  - Minimum HR for statistical significance: 0.785

Interim Analysis

- Important aspect of design: allows early stopping for efficacy and futility
- Interim analysis are event driven: similar to primary
- Very similar methodology as for non-TTE endpoints
- Group sequential design is the gold standrad
  - α-spending and β- spendings are widely used
  - Other methods are available
- Efficacy and futility boundaries are “non-binding”

Design Challenges with Potential NPH

- Potential of NPH brings more uncertainties in design assumption
- Treatment differences under NPH constitute a broad class of alternative hypotheses
  - Degree of effect
  - Delayed timing of effect: Delayed separation of survival curves
  - Different effects in unanticipated subpopulations: Can result in crossing hazards
  - Diminishing effect over time
- How do we design a trial to be powerful across MANY alternatives?

Impact of NPH on Traditional Design

- Consider the alternative of delay in treatment effect: 2 months

$$\begin{aligned} HR &= 1 \qquad t \leq 2 months \\ &= 0.67 \quad t > 2 months \end{aligned}$$

- With 263 events and 300 patients
  - **Power = 63%** $\downarrow$
- Need 520 events and 600 patients for power= $\sim$ 90%
- Significant increase in resources: Sample size doubled
  - Standard log-rank test based failed to show robust power under different alternatives

Design Under NPH: General Considerations

- Trial duration or total follow up time plays an important role
  - Event based only analysis may produce a design that finishes too early Underpowered
  - May fail to describe time dependent treatment effect
- Carefully elicitation of the possible treatment effect scenario
  - Power trial for multiple scenarios
  - Find worst-case scenario, e.g.,
  - Minimum effect size of interest (PH)
  - Delayed effect
  - Early crossing hazards

Interim Analysi Under NPH: General Considerations

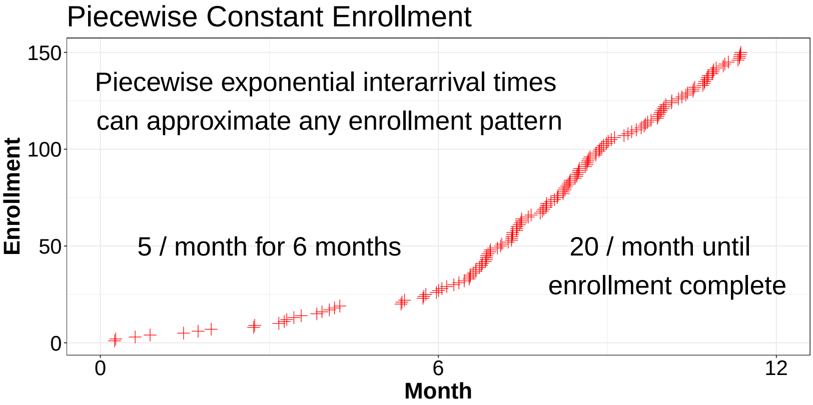
- Need careful consideration
  - Especially, for late emerging treatment effect scenarios
  - An early interim analysis will have smaller probability of stopping for efficacy and higher probability of crossing any futility bound
- Balance between the risks of stopping too soon before late benefit emerges and the appropriately monitoring of the trial for futility
- Futility analysis: is it really necessary?
  - Safety bound or conditional power based approaches can be useful
- MaxCombo requires set timing based on events AND follow-up to ensure power

6. Practical Designs in Presence of Non-proportional Hazard

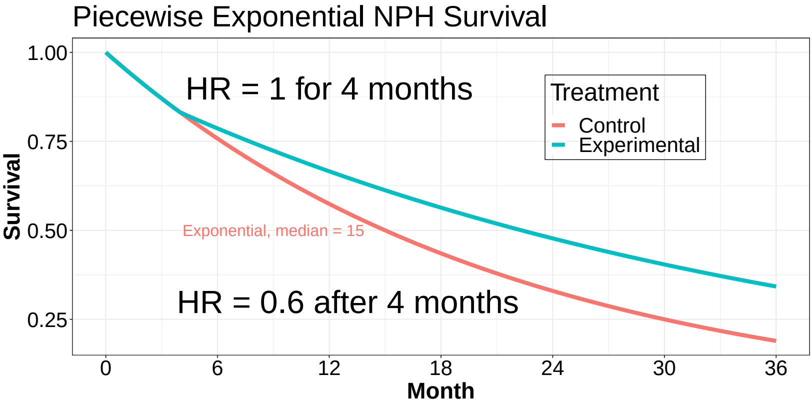
Introducting the Piecewise Model

- Simple model to approximate arbitrary patterns of
  - Enrollment: piecewise constant enrollment rates
  - Failure rates: piecewise exponential
  - Dropout rates: piecewise exponential
- Combined tools for designing and evaluating designs
  - Asymptotic approach using average hazard ratio (AHR)
  - Simulation tools to confirm asymptotic approximations
  - No requirement for proportional hazards
  - Stick with logrank for today

Piecewise Constant Enrollment



Simple NPH Example



Average Hazard Ratio (AHR)

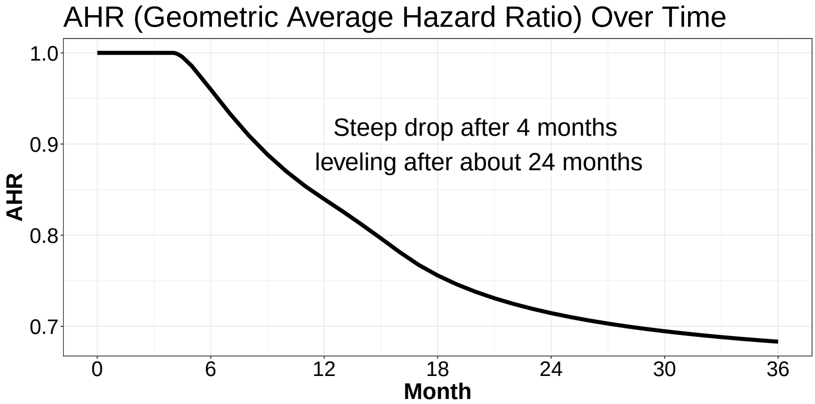
- Geometric mean hazard ratio (Mukhopadhyay et al. (2020))
- Exponentiate: average log(HR) weighted by expected events per interval

Interval	HR	-ln(HR)	Expected Events
0-4	1.0	0.00	d1
>4	0.6	0.51	d2

$$\text{AHR} = \exp\left(\frac{d_1 \log(1) + d_2 \log(0.6)}{d_1 + d_2}\right)$$

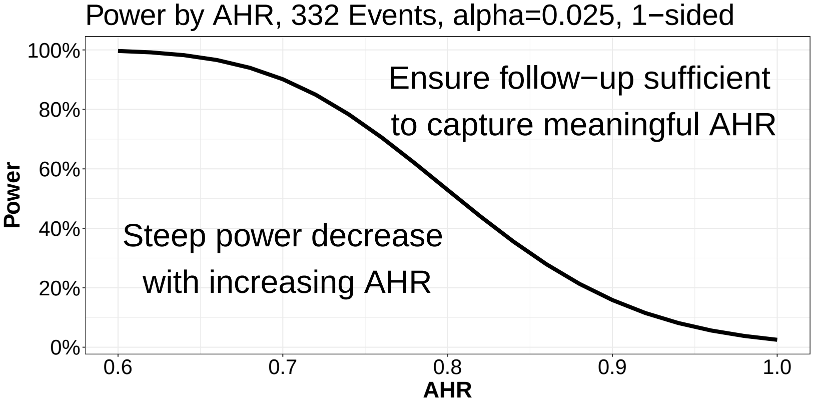
AHR Over Time

- Constant enrollment rate, 12 month targeted enrollment
- Exponential dropout, 0.001 per month
- Control: exponential, median = 15 months
- HR: 1 in months 0-4, 0.6 thereafter



Power by AHR

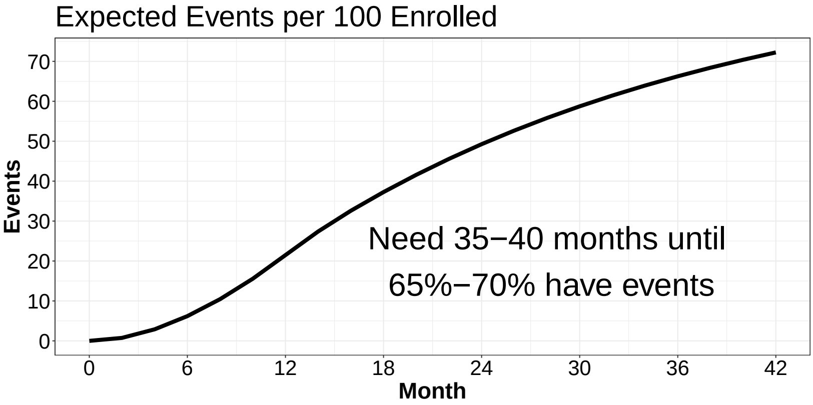
Assume 332 events



AHR as Estimand

- Some argue this is a bad idea
    - e.g., hazards of hazard ratios (Hernán (2010))
- Pro's
    - Estimated by Cox regression
    - AHR concept makes more clear what this is
    - Logrank is widely-accepted corresponding test
    - Stable target if follow-up sufficient
    - Both asymptotic approximations and simulation supported (today!)
        - This includes group sequential design
    - Easy to approximate arbitrary enrollment, failure and dropout patterns
- Cautions
    - No single estimand sufficently describes NPH differences
    - Early interim analysis (futility, efficacy) should anticipate possible reduced effect

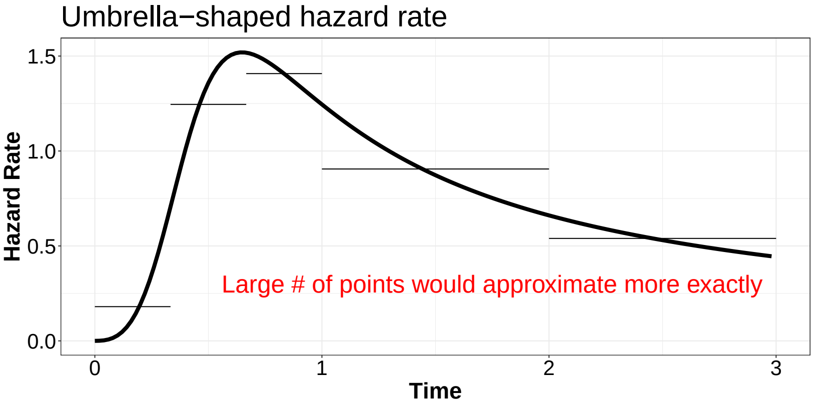
Expected Accrual of Endpoints



Expected Events per 100 Enrolled

Need 35−40 months until 65%−70% have events

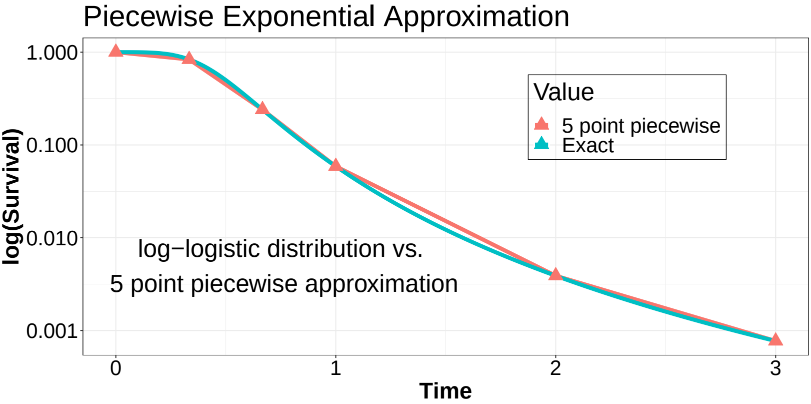
Piecewise Exponential Approximation of Log-Logistic



Large # of points would approximate more exactly

Piecewise Exponential Approximation of Log-Logistic

Approximate any survival distribution



More points = better approximation

Asymptotic Approximation

- Use of Tsiatis (1982) (also extends to weighted logrank; not discussed today)
- Statistical information proportional to expected event counts as in Schoenfeld (1981)
- Natural parameter: log(AHR)
- Statistical information and test correlation still $\sim$proportional to number of events
- Extension of Jennison and Turnbull (2000) calculations to non-constant effect size over time
- Subject of forthcoming paper

Asymptotic Distribution Simplified

Statistical information at analysis $1 \leq k = \leq K$: I_k

Proportion of final information at analysis k: $t_k = I_k / I_K$

$$Z_k \sim \text{Normal}(\sqrt{I_k}\theta(t_k), 1)$$

Multivariate normal with correlations for $1 \leq j \leq k \leq K$:

$$\text{Corr}(Z_j, Z_k) = \sqrt{t_j / t_k}$$

Asymptotic Boundary Crossing Probabilities

- Note dependence on time-varying $\theta(t_k)$, $1 \leq k \leq K$
- θ notation below does not explicitly clarify changing values with time
- Bounds $-\infty \leq a_k < b_k \leq \infty$ for $1 \leq k < K$, $-\infty \leq a_K \leq b_K < \infty$
- Upper boundary crossing probabilities

$$\alpha_k(\theta) = P_\theta(\{Z_k \geq b_k\} \cap_{j=1}^{i-1} \{a_j \leq Z_j < b_j\})$$

- Lower boundary crossing probabilities

$$\beta_k(\theta) = P_\theta((Z_k < a_k\} \cap_{j=1}^{k-1} \{a_j \leq Z_j < b_j\}).$$

- Boundary crossing probabilities computed with simple extension of Jennison and Turnbull (2000) algorithm
  - For now, you can cite **gsdmvn** R package at GitHub

Spending bounds

- Spending bounds also computed with simple extension of Jennison and Turnbull (2000) algorithms
- For lower bound, lesser early treatment effect is accounted for!

7. Implementation using R- Part II

Simulation Tools: simtrial Package

- Low-level tools to demonstrate model
- Higher-level tools to enable trial simulations
  - Fixed designs
  - Group sequential designs

simtrial: lower-level routines

We will not go into these today

- `fixedBlockRand()` - fixed block randomization
- `rpwenroll()` - random inter-arrival times with piecewise constant enrollment rates
- `rpwexp()` - piecewise exponential failure rate generation
- `cutData()` - cut data for analysis at a specified calendar time
- `cutDataAtCount()` - cut data for analysis at a specified event count, including ties on the cutoff date
- `getCutDateForCount()` - find date at which an event count is reached
- `tensurv()` - pre-process survival data into a counting process format

Generating a trial

Stratification and blocking (used for simulation; not needed for design)

```r
# 2 strata
strata <- tibble(Stratum=c("All"))

# Block size of 4, equal randomization; VECTOR ARGUMENT
block <- c(rep("Control",2),rep("Experimental",2))
```

Enrollment rates

```r
# 1 year enrollment, increasing rates
enrollRates <-
  tibble(Stratum = "All", duration = 12, rate = 476 / 12)
```

Generating a trial

Failure rates

```r
# Control: exponential with 15 month median
# HR: 1 for 4 months, 0.6 thereafter
failRates <- tribble(
  ~Stratum, ~duration, ~failRate,   ~hr, ~dropoutRate,
  "All",    4,         log(2) / 15,  1,   .001,
  "All",    100,       log(2) / 15,  0.6, .001)
```

Generate a Trial

Simple simulation; fixed design

```
sim <-
simtrial::simfix(nsim=50, sampleSize=476, targetEvents=332,
  strata, enrollRates, failRates, totalDuration=36, block,
  timingType=1:5) %>% mutate(AHR = exp(lnhr))
head(sim, n=5) %>% kable(digits=2) %>%
  kable_styling(font_size=8)
```

Events	lnhr	Z	cut	Duration	Sim	AHR
321	-0.35	-3.09	Planned duration	36.00	1	0.71
332	-0.33	-3.04	Targeted events	37.80	1	0.72
322	-0.34	-3.05	Minimum follow-up	36.55	1	0.71
332	-0.33	-3.04	Max(planned duration, event cut)	37.80	1	0.72
332	-0.33	-3.04	Max(min follow-up, event cut)	37.80	1	0.72

Trial Simulation: MaxCombo

MaxCombo test set up

```r
# Set up tests to be used
rg <- tibble(rho=c(0,0,.5), gamma=c(0,.5,.5))
rg %>% kable()
```

rho	gamma
0.0	0.0
0.0	0.5
0.5	0.5

Simulating Multiple Tests for MaxCombo

```
sim <-
simtrial::simfix(nsim=50, sampleSize=476, targetEvents=332,
  strata, enrollRates, failRates, totalDuration=36, block,
  timingType=2, rg = rg) %>%
  select(c(Sim, Events, Duration, rho, gamma, Z, V1, V2, V3))
```

Simulation Output

```
sim %>%
head(sim, n=6) %>%
  kable(digits=c(0,2,1,1,2,2,2,2,2)) %>%
  kable_styling(font_size=8)
```

Sim	Events	Duration	rho	gamma	Z	V1	V2	V3
1	332	41.3	0.0	0.0	-2.90	1.00	0.94	0.97
1	332	41.3	0.0	0.5	-3.31	0.94	1.00	0.99
1	332	41.3	0.5	0.5	-3.23	0.97	0.99	1.00
2	332	41.9	0.0	0.0	-3.62	1.00	0.94	0.97
2	332	41.9	0.0	0.5	-4.16	0.94	1.00	0.99
2	332	41.9	0.5	0.5	-4.02	0.97	0.99	1.00

Generate a Trial: Power Estimates

Summarize simulations by weighting scheme

```r
sim %>% group_by(rho,gamma) %>%
  summarize(Power=mean(Z<=qnorm(.025)),
            Duration=mean(Duration),
            Simulations = n()) %>%
  kable(digits=3) %>% kable_styling(font_size=8)
```

rho	gamma	Power	Duration	Simulations
0.0	0.0	0.90	38.935	50
0.0	0.5	0.96	38.935	50
0.5	0.5	0.96	38.935	50

Note weighted logrank improvements over logrank

Summarize MaxCombo

Power estimated using `pMaxCombo()` function for each simulation

```r
# subset to targeted events cutoff tests
p <- unlist(sim %>% group_by(Sim) %>% group_map(pMaxCombo))
mean(p<.025)
```

```
## [1] 0.96
```

MaxCombo also has higher power than logrank

Generate a Trial: Group Sequential

Generating a trial step-by-step allows more flexibility

```r
a <- simfix2simPWSurv(failRates)
x <- simPWSurv(n = 400, # Sample size
               strata = strata,
               block = block,
               enrollRates = enrollRates,
               failRates = a$failRates,
               dropoutRates = a$dropoutRates)
```

Generate a Trial

Resulting format

Stratum	enrollTime	Treatment	failTime	dropoutTime	cte	fail
All	0.044	Control	39.184	57.173	39.228	1
All	0.073	Experimental	14.561	2905.279	14.634	1
All	0.077	Control	26.369	560.024	26.446	1
All	0.147	Experimental	40.538	1386.317	40.685	1
All	0.164	Control	27.242	817.336	27.405	1
All	0.168	Experimental	17.691	188.886	17.858	1

Simulate Repeatedly

Repeated simulations analyzed after 150 and 250 events

```
y <- NULL
for(sim in 1:3){
x <- simPWSurv(n = 400, # Sample size
               strata = strata,
               block = block,
               enrollRates = enrollRates,
               failRates = a$failRates,
               dropoutRates = a$dropoutRates)
for(Events in c(150,250)){
y <- rbind(y, x %>% cutDataAtCount(Events) %>%
             tensurv(txval="Experimental")%>%
             tenFH(rg=tibble(rho=0,gamma=0)) %>%
             mutate(sim=sim, Events=Events))
}}
```

Simulate Repeatedly

rho	gamma	Z	sim	Events
0	0	-2.287	1	150
0	0	-3.990	1	250
0	0	-1.563	2	150
0	0	-2.390	2	250
0	0	-2.565	3	150
0	0	-3.971	3	250

AHR Tools: gsDesign2 package

Main functions of interest today under piecewise model:

- `s2pwe()`: approximate arbitrary survival distribution with piecewise exponential
- `eEvents_df()`: expected event accrual over time
- `AHR()`: average hazard ratio over time

Approximating Using Piecewise Model

Approximating log-logistic distribution plotted above using piecewise model

```r
dloglogis <- function(x, alpha = 1, beta = 4){
  1 / (1 + (x/alpha)^beta)
}
times10 <- c(seq(1/3,1,1/3),2,3)
# Use s2pwe() to generate piecewise approximation
s2pwe(times10,dloglogis(times10,alpha=.5,beta=4)) %>%
  kable(digits=3)
```

duration	rate
0.333	0.541
0.333	3.736
0.333	4.223
1.000	2.716
1.000	1.619

Approximating Event Accumulation Over Time

This basic calculation is driving much of what we do today!

```r
gsDesign2::eEvents_df(enrollRates,
               failRates=tibble(duration=c(3,100),
                                failRate = c(.1,.05),
                                dropoutRate = .001),
  totalDuration = 23)
```

```
## [1] 296.4448
```

Approximating AHR Over Time

This basic calculation is driving much of what we do today!

```
gsDesign2::AHR(enrollRates, failRates,
  totalDuration = seq(12,36,12)) %>%
  kable(digits=c(0,2,0,1,1))
```

Time	AHR	Events	info	info0
12	0.84	102	25.1	25.6
24	0.71	234	57.2	58.6
36	0.68	315	77.5	78.8

Group Sequential Design Tools: gsdmvn package

Main functions of interest today:

- `gs_design_ahr()`: design under non-proportional hazards
- `gs_power_ahr()`: power under non-propotional hazards
- `gs_spending_bound()`: spending bound specification
- `gs_b()`: Fixed boundary generation

Fixed design

Set up libraries and rate assumptions

```
library(tibble)
library(gsdmvn)
library(dplyr)
library(knitr)
enrollRates <- tibble(Stratum="All", duration = 12, rate = 1)
failRates <- tibble(Stratum="All",
                    duration=c(4, 100),
                    failRate = log(2)/ 15,
                    hr = c(1, .6),
                    dropoutRate = 0.001)
```

Fixed design

```
# Single analysis
x <-
gs_design_ahr(enrollRates,
              failRates,
              analysisTimes = 36, # Single analysis
              upper=gs_b, upar = qnorm(.975), # Z for p=.025
              lower=gs_b, lpar = -Inf) # No lower bound
x$bounds %>% filter(Bound == "Upper") %>%
  select(-c(Analysis,Bound)) %>%
  kable(digits=c(0,0,0,2,2,2,2,2,2)) %>%
  kable_styling(font_size = 8)
```

Time	N	Events	Z	Probability	AHR	theta	info	info0
36	440	292	1.96	0.9	0.68	0.38	71.63	72.9

Round up sample size and events (not done here!)

Group Sequential Design

- Spending function for upper bound

```r
# upper is a function to compute bound
upper <- gs_spending_bound
# upar is a parameter passed to upper
upar <- list(sf = gsDesign::sfLDOF,
             total_spend = 0.025,
             param = NULL)
```

- `sf`: spending function from gsDesign package
- `total_spend`: for upper bound, this is α-spending
- `param`: parameter to pass to spending function, if needed

Group Sequential Design

- Spending function for lower bound

```r
# lower is a function to compute bound
lower <- gs_spending_bound
# lpar is a parameter passed to upper
lpar <- list(sf = gsDesign::sfHSD,
             total_spend = 0.1,
             param = -2)
```

- `sf`: in this case, Hwang-Shih-DeCani spending function
- `total_spend`: in this case, Type II or β-spending (90% power $= 100(1-\beta)$)
- `param`: in this case, `param=2` is passed to `sfHSD()` to realize $\gamma = 2$
- Lan-DeMets spending function to approximate O'Brien-Fleming bound

Group Sequential Spending Function Design

```
x <- gs_design_ahr(enrollRates, failRates,
          alpha=0.025, beta=.1,
          # information fraction at analyses
          IF = c(.67, .85, 1),
          # total planned trial duration
          analysisTimes = 36,
          # Spending bounds as before
          upper=upper, upar=upar,
          lower=lower, lpar=lpar)
names(x)
```

```
## [1] "enrollRates" "failRates"   "bounds"
```

Enrollment Rates Required

```r
# Enrollment rates for design
x$enrollRates %>% kable(digits=2)
```

Stratum	duration	rate
All	12	41.62

Design Bounds

```r
x$bounds %>% select(-c(theta, info, info0)) %>%
  # Round up sample size and event count
  mutate(N=ceiling(N/2) * 2,
         Events = ceiling(Events)) %>%
  kable(digits=c(0,0,1,0,0,2,2,2)) %>%
  kable_styling(font_size = 8)
```

Analysis	Bound	Time	N	Events	Z	Probability	AHR
1	Upper	21.4	500	222	2.50	0.43	0.73
2	Upper	28.3	500	282	2.22	0.77	0.70
3	Upper	36.0	500	331	2.05	0.90	0.68
1	Lower	21.4	500	222	0.62	0.04	0.73
2	Lower	28.3	500	282	1.38	0.07	0.70
3	Lower	36.0	500	331	2.05	0.10	0.68

Simulation to Confirm Design Properties

- Easiest way to confirm that asymptotic approximation works
- We will demonstrate the steps required for this
- Final simulation will look both at logrank and MaxCombo test
  - MaxCombo will improve power and control Type I error
  - Will not go through full detail on deriving final MaxCombo bound
- Will use relatively detailed code here as there are lots of options
  - You may wish to write a function to simplify

Trial Generation and Analysis

We demonstrate a single trial simulation

```r
fr <- simfix2simPWSurv(failRates)
N <- ceiling(max(x$bound$N))
# Generate a single trial
d <- simPWSurv(n=N, enrollRates = enrollRates,
               failRates = fr$failRates,
               dropoutRates = fr$dropoutRates)
# Get event count planned at each analyss
ev <- ceiling(sort(unique(x$bounds$Events)))
# Set place to save analyses
y <- NULL
mc <- NULL
# Set up rho, gamma combinations for MaxCombo
# logrank, (0, .5), (.5, .5)
rg <- tibble(rho=c(0,0,.5), gamma=c(0,.5,.5))
```

Do Interim and Final Analyses

Loop through analyses and accumulate results

```r
for(Analysis in seq(ev)){
  # Cut data for analysis and get counting process format
  a <- d %>% cutDataAtCount(ev[Analysis]) %>%
       tensurv(txval="Experimental")
  # Do logrank
  y <- rbind(y, a  %>% tenFH(rg=tibble(rho=0, gamma=0)) %>%
               mutate(Analysis = Analysis))
  # At final analysis, compute MaxCombo
  if(Analysis == length(ev)){
    mc <- rbind(mc, a %>% tenFHcorr(rg))
  }}
```

Interim and Final Analysis Results

Logrank

```
y %>% kable(digits=2)
```

rho	gamma	Z	Analysis
0	0	-2.19	1
0	0	-3.32	2
0	0	-3.69	3

Component Tests of MaxCombo at Final Analysis

Weighted logrank Z-test often larger than logrank under delayed effect

```
mc %>% kable(digits=2)
```

rho	gamma	Z	V1	V2	V3
0.0	0.0	-3.69	1.00	0.94	0.97
0.0	0.5	-4.37	0.94	1.00	0.96
0.5	0.5	-3.92	0.97	0.96	1.00

MaxCombo p-value (single analysis)

```
pMaxCombo(mc)
```

```
## [1] 6.079523e-06
```

Comparing Simulation Results to Group Sequential Bounds

Reformat bounds

```
b <-
x$bounds %>% select(c(Analysis, Bound, Z)) %>%
  tidyr::pivot_wider(names_from="Bound", values_from="Z")
b %>% kable(digits=2)
```

Analysis	Upper	Lower
1	2.50	0.62
2	2.22	1.38
3	2.05	2.05

Combine Bounds with Analyses

```r
left_join(y, b, by = "Analysis") %>% kable(digits=2)
```

rho	gamma	Z	Analysis	Upper	Lower
0	0	-2.19	1	2.50	0.62
0	0	-3.32	2	2.22	1.38
0	0	-3.69	3	2.05	2.05

Select Critical Analysis

Select first analysis with bound crossed or final analysis

```r
left_join(y, b, by = "Analysis") %>%
  mutate(Stop=(-Z>=Upper | -Z < Lower | Analysis == 4)) %>%
  filter(Stop == TRUE) %>% slice(1) %>% kable(digits=3)
```

rho	gamma	Z	Analysis	Upper	Lower	Stop
0	0	-3.322	2	2.22	1.378	TRUE

- Now we see where the critical analysis was for this simulation.
- Do this repeatedly and you can summarize the group sequential properties of design.

Group Sequential Design: Asymmetric Design Example

- Use a fixed lower bound
  - Futility only for safety at IA 1 (e.g., p=0.05 in wrong direction)
  - No futility bound after IA 1
  - Non-binding futility bound is default
- Spending bound for upper bound
  - O'Brien-Fleming often urged by regulators
  - As before; no efficacy test at early safety analysis

```r
lower <- gs_b # Fixed lower bound
# Futility testing only at early analysis
lpar <- c(qnorm(.05), rep(-Inf, 3))
# Efficacy testing only AFTER first analysis
test_upper <- c(FALSE, rep(TRUE, 3))
# Timing now set based on trial duration
analysisTimes <- c(12,20,28,36)
```

Group Sequential Spending Function Design

```
x <- gs_design_ahr(enrollRates, failRates,
        alpha=0.025, beta=.1,
        # calendar timing of all analyses
        analysisTimes = analysisTimes,
        # bounds from previous slides
        upper=upper, upar=upar,
        lower=lower, lpar=lpar,
        test_upper=test_upper)
```

Bounds for Group Sequential Design

```
x$bounds %>% select(-c(theta,info,info0)) %>%
  mutate(N=ceiling(N/2)*2,
         Events=ceiling(Events)) %>%
  kable(digits=c(rep(0,5),rep(2,3))) %>%
  kable_styling(font_size = 8)
```

Analysis	Bound	Time	N	Events	Z	Probability	AHR
1	Upper	12	468	101	Inf	0.00	0.84
2	Upper	20	468	195	2.60	0.31	0.74
3	Upper	28	468	262	2.22	0.74	0.70
4	Upper	36	468	310	2.05	0.90	0.68
1	Lower	12	468	101	-1.64	0.01	0.84
2	Lower	20	468	195	-Inf	0.01	0.74
3	Lower	28	468	262	-Inf	0.01	0.70
4	Lower	36	468	310	-Inf	0.01	0.68

Symmetric Design

- In this case h1_spending=FALSE indicates lower spending under null hypothesis

```
x <- gs_design_ahr(enrollRates, failRates,
        alpha=0.025, beta=.1,
        # For symmetric design, use binding bounds
        binding = TRUE,
        # calendar timing of all analyses
        analysisTimes = c(20, 28, 36),
        upper=upper, upar=upar,
        # copied upper bound to lower bound
        lower=upper, lpar=upar,
        # use this for symmetric bound
        h1_spending=FALSE)
```

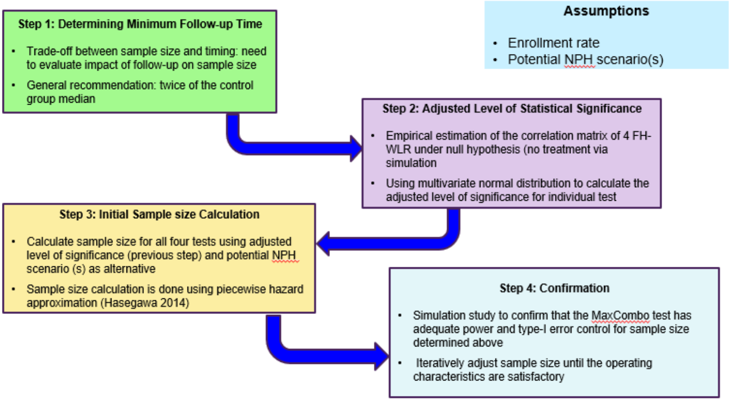
Bounds for Symmetric Design

```
x$bounds %>%
  mutate(N=ceiling(N/2)*2,
         Events=ceiling(Events)) %>%
  select(-c(theta,info,info0)) %>%
  kable(digits=c(rep(0,5),rep(2,3))) %>%
  kable_styling(font_size = 8)
```

Analysis	Bound	Time	N	Events	Z	Probability	AHR
1	Upper	20	466	194	2.60	0.30	0.74
2	Upper	28	466	260	2.22	0.73	0.70
3	Upper	36	466	309	2.05	0.90	0.68
1	Lower	20	466	194	-2.60	0.00	0.74
2	Lower	28	466	260	-2.22	0.00	0.70
3	Lower	36	466	309	-2.05	0.00	0.68

8. Designing TTE Trial with MaxCombo Test

Sample Size Calculation: Two Step Approach (Roychoudhury et al. (2020))



Assumptions

- Enrollment rate
- Potential NPH scenario(s)

Step 1: Determining Minimum Follow-up Time

- Trade-off between sample size and timing: need to evaluate impact of follow-up on sample size
- General recommendation: twice of the control group median

Step 2: Adjusted Level of Statistical Significance

- Empirical estimation of the correlation matrix of 4 FH-WLR under null hypothesis (no treatment via simulation
- Using multivariate normal distribution to calculate the adjusted level of significance for individual test

Step 3: Initial Sample size Calculation

- Calculate sample size for all four tests using adjusted level of significance (previous step) and potential NPH scenario (s) as alternative
- Sample size calculation is done using piecewise hazard approximation (Hasegawa 2014)

Step 4: Confirmation

- Simulation study to confirm that the MaxCombo test has adequate power and type-I error control for sample size determined above
- Iteratively adjust sample size until the operating characteristics are satisfactory

Group Sequential Design with MaxCombo Test

- Use of log-rank test for interim analysis and MaxCombo for final analysis
  - To avoid the impact of shorter follow up time or trial duration in WLR
  - Well accepted by the regulators
- Final success boundary needs multiplicity adjustment due to the correlation between the LR
  test at interim and the MaxCombo test in final analysis
- We propose calculation of the final boundary using independent increment of information from interim to final and asymptotic normality
- The impact on type I error and power for interim analysis need to be evaluated via simulation

9. Summary and Discussion

Summary - I

- LR test and Cox regression are still gold standard
- Use MaxCombo or modified MaxCombo for primary statistical testing: a combination test based on FH-WLR tests
  - Extensive simulation study shows better statistical power of the MaxCombo test over traditional LRT under various types of NPH (especially for delayed treatment effect)
  - Maintains good statistical properties under PH
- No single statistical measure can capture the time dependent nature of treatment benefit
  - Proposed stepwise approach provides a complete summary

Summary - II

- Under potential NPH, design should specify sample size and total follow-up time to ensure adequate power and type-I error
- Piecewise exponential approximation provides a flexible way to design TTE trial under NPH
- Efficient R packages are critical to implement the non-traditional design in real-life: simtrial, gsdesign2
    - Simulation plays an important role
- Design and analysis be pre-specified in the protocol and SAP to comply with ICH-E9

Cross-Industry Working Group

- NPH working group (WG) focused on statistical methods
  - Beyond LRT and Cox regression model in presence of NPH
  - Can be pre-specified in the statistical analysis plan (SAP)
  - Aids with interpretation of treatment benefit
- First meeting in October 2016: ASA Regulatory-Industry workshop
  - Face to Face mid-point meeting June 2017: ASCO
  - Presentation of key findings and February 2018: Duke-Margolis Workshop
  - Face to Face November 2019

References

References I

Anderson, Keaven M. 2020a. “Clinical Trial Simulation.” *GitHub Repository*. https://github.com/keaven/simtrial; Merck & Co., Inc.

———. 2020b. “Group Sequential Design Under Non-Proportional Hazards.” *GitHub Repository*. https://github.com/keaven/gsDesign2; Merck & Co., Inc.

———. 2020c. “Group Sequential Design with Non-Constant Effect.” *GitHub Repository*. https://github.com/keaven/gsdmvn; Merck & Co., Inc.

Breslow, N. E., L. Edler, and J. Berger. 1984. “A Two-Sample Censored-Data Rank Test for Acceleration.” *Biometrics* 40 (4): 1049–62.

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

References II

Callegaro, Andrea, and Bart Spiessens. 2017. "Testing Treatment Effect in Randomized Clinical Trials with Possible Nonproportional Hazards." *Statistics in Biopharmaceutical Research* 9 (2): 204–11.

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistics Society, Series B* 34: 187–220.

Freidlin, Boris, and Edward L. Korn. 2019. "Methods for Accommodating Nonproportional Hazards in Clinical Trials: Ready for the Primary Analysis?" *Journal of Clinical Oncology* 37 (35): 3455–9.

Genz, A. 1992. "Numerical Computation of Multivariate Normal Probabilities." *Journal of Computational and Graphical Statistics* 1 (2): 141–49.

Grambsch, Patricia M., and Terry M. Therneau. 1994. "Proportional Hazards Tests and Diagnostics Based on Weighted Residuals." *Biometrika* 81 (3): 515–26.

References III

Hernán, Miguel A. 2010. "The Hazards of Hazard Ratios." *Epidemiology (Cambridge, Mass.)* 21 (1): 13.

Jennison, Christopher, and Bruce W. Turnbull. 2000. *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, FL: Chapman; Hall/CRC.

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

Klein, John P., Brent Logan, Mette Harhoff, and others. 2007. "Analyzing Survival Curves at a Fixed Point in Time." *Statistics in Medicine* 26 (24): 4505–19.

Lee, Seung-Hwan. 2007. "On the Versatility of the Combination of the Weighted Log-Rank Statistics." *Computational Statistics and Data Analysis* 51 (12): 6557–64.

References IV

Lin, Ray S., Ji Lin, Satrajit Roychoudhury, and others. 2020. “Alternative Analysis Methods for Time to Event Endpoints Under Nonproportional Hazards: A Comparative Analysis.” *Statistics in Biopharmaceutical Research* 12 (2): 187–98.

Logan, Brent R., John P. Klein, and Mei-Jie Zhang. 2008. “Comparing Treatments in the Presence of Crossing Survival Curves: An Application to Bone Marrow Transplantation.” *Biometrics* 64 (3): 733–40.

Magirr, Dominic, and Carl-Fredrik Burman. 2019. “Modestly Weighted Logrank Tests.” *Statistics in Medicine* 38 (20): 3782–90.

Moore, Malcolm J., David Goldstein, John Hamm, and others. 2007. “Erlotinib Plus Gemcitabine Compared with Gemcitabine Alone in Patients with Advanced Pancreatic Cancer: A Phase Iii Trial of the National Cancer Institute of Canada Clinical Trials Group.” *Journal of Clinical Oncology* 25 (15): 1960–6.

References V

Mukhopadhyay, Pralay, Wenmei Huang, Paul Metcalfe, Fredrik Öhrn, Mary Jenner, and Andrew Stone. 2020. "Statistical and Practical Considerations in Designing of Immuno-Oncology Trials." *Journal of Biopharmaceutical Statistics*. https://doi.org/10.1080/10543406.2020.1815035.

Pepe, Margaret Sullivan, and Thomas R. Fleming. 1989. "Weighted Kaplan-Meier Statistics: A Class of Distance Tests for Censored Survival Data." *Biometrics* 45 (2): 497–507.

Perón, Julien, Pascal Roy, Brice Ozenne, and others. 2016. "The Net Chance of a Longer Survival as a Patient-Oriented Measure of Treatment Benefit in Randomized Clinical Trials Measurement of the Net Chance of a Longer SurvivalMeasurement of the Net Chance of a Longer Survival." *JAMA Oncology* 2 (7): 901–5.

Putter, H., M. Sasako, H. H. Hartgrink, and others. 2005. "Long-Term Survival with Non-Proportional Hazards: Results from the Dutch Gastric Cancer Trial." *Statistics in Medicine* 24 (18): 2807–21.

References VI

Roychoudhury, Satrajit, Keaven M Anderson, Jiabu Ye, and Pralay Mukhopadhyay. 2020. "Robust Design and Analysis of Clinical Trials with Non-Proportional Hazards: A Straw Man Guidance from a Cross-Pharma Working Group." https://arxiv.org/abs/1908.07112.

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): 2409–21.

Schemper, Michael, Samo Wakounig, and Georg Heinze. 2009. "The Estimation of Average Hazard Ratios by Weighted Cox Regression." *Statistics in Medicine* 28 (19): 2473–89.

Schoenfeld, David. 1981. "The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions." *Biometrika* 68 (1): 316–19.

References VII

Tsiatis, Anastasios A. 1982. "Repeated Significance Testing for a General Class of Statistics Use in Censored Survival Analysis." *Journal of the American Statistical Association* 77: 855–61.

Uno, Hajime, Brian Claggett, Lu Tian, and others. 2014. "Moving Beyond the Hazard Ratio in Quantifying the Between-Group Difference in Survival Analysis." *Journal of Clinical Oncology* 32 (22): 2380–5.

Xu, Zhenzhen, Yongsoek Park, Boguang Zhen, and others. 2018. "Designing Cancer Immunotherapy Trials with Random Treatment Time-Lag Effect." *Statistics in Medicine* 37 (30): 4589–4609.

Xu, Zhenzhen, Boguang Zhen, Yongsoek Park, and others. 2017. "Designing Therapeutic Cancer Vaccine Trials with Delayed Treatment Effect." *Statistics in Medicine* 36 (4): 592–605.

Yang, Song, and Ross Prentice. 2010. "Improved Logrank-Type Tests for Survival Data Using Adaptive Weights." *Biometrics* 66 (1): 30–38.

Design for the MaxCombo Test Under Non-Proportional Hazards

Contents

```r
defaultW <- getOption("warn")
options(warn = -1) # Suppress loading messages
source("SS_HF.r")
library(gsDesign)
library(simtrial)
library(survival)
library(dplyr)
library(kableExtra)
library(mvtnorm)
options(warn = defaultW)
```

1 Overview

1.1 Outline of the proposed method

This document suggests methods for designing a study when there is a reasonable possibility that the hazard ratio for the two treatment groups will not be constant over time. The suggestion is to ensure power for some 'worst case' non-proportional hazards situation and a proportional hazards case with a lesser late benefit but the same sample size and event count required for powering. Because of the non-proportional hazards possibility, we assume testing will be done with the maximum of 4 weighted logrank (Fleming-Harrington) tests: G(0,0) which is the logrank test, G(0,1) which down-weights early events, G(1,0) which down-weights late events, and G(1,1) which down-weights early and late events relative to those in the middle of the distribution. The design strategy is to assume an enrollment duration and then vary the assumed minimum follow-up after enrollment to optimize perceived tradeoffs between sample size and trial duration. With this enrollment and follow-up pattern, we simulate the null hypothesis to approximate correlation between the components of the MaxCombo test and then apply the multivariate normal distribution to compute an adjusted nominal alpha level to test at to ensure 2.5% one-sided Type I error; this is based on results by Karrison and others (2016). Next, we apply the method of Hasegawa (2014) to approximate the sample size and event count required to power the design for each component of the MaxCombo test; a minor modification of the Hasegawa method allows crossing rather than just delayed treatment effect. The minimum of these is selected for the design sample size and power is verified by simulation. We then consider extending this to a case with a single interim analysis to demonstrate use of the multivariate normal distribution to ensure multiplity control across analyses and all components of the MaxCombo test; for simplicity and to ensure possible better regulatory acceptance of a positive interim finding, we assume interim testing will be done with only the logrank test.

1.2 Installing packages

The primary tool used here is the **simtrial** package. This is a minimal package written using tidy coding and attempting to allow validation to enable use for regulatory submissions. As of the current release, validation is continuing; however, independent double programming using the **nphsim** package has been used to check the weighted logrank and MaxCombo computations. Installation of **simtrial** from GitHub is performed as follows:

```
devtools::install_github('keaven/simtrial')
```

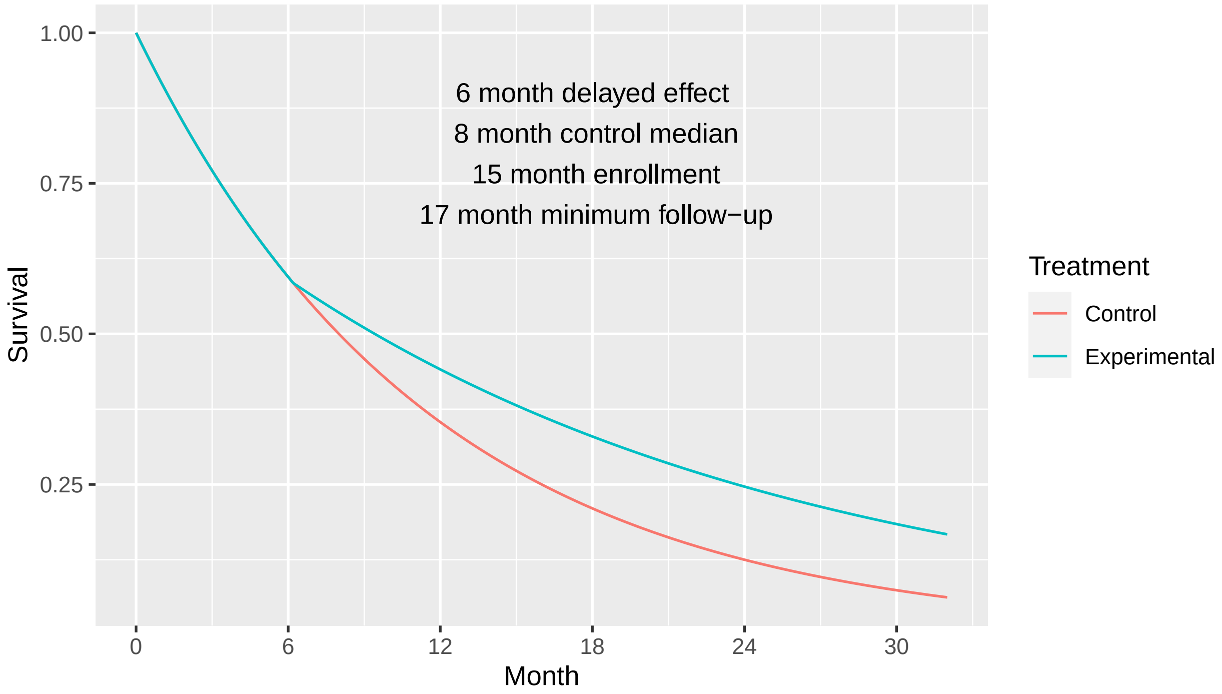
Help files and vignettes available with the **simtrial** should be helpful for further clarification of questions that may arise concerning code.

2 Initial considerations

2.1 Assumptions for the example implemented

Throughout this document we will assume 15 months of constant enrollment, a constant dropout rate of 0.001 per month, control group observations follow an exponential survival distribution with a median of 8 months, no treatment effect for 6 months (HR=1), followed by a hazard ratio of 0.56 thereafter.

```
de<-rbind(tibble(Treatment="Control",Month=(0:(32*5))/5,lambda=log(2)/8),
          tibble(Treatment="Experimental",Month=(0:(32*5))/5) %>%
                 mutate(lambda=log(2)/8*ifelse(Month<=6,1,.56))) %>%
          group_by(Treatment) %>%
          mutate(h=exp(-lambda/5),Survival=lag(cumprod(h),default = 1))
lbl <- "6 month delayed effect\n 8 month control median\n 15 month enrollment\n 17 month minimum follow-
print(ggplot(de,aes(x=Month,y=Survival,col=Treatment)) +
      geom_line() +
      annotate(geom="text", x=18, y=.8, label=lbl,color="black",size=4) +
      scale_x_continuous(breaks=(0:5)*6) +
      scale_y_continuous())
```



2.2 The sample size time tradeoff

We consider total trial duration of 18, 24, 32 and 40 months to compare required sample size each component of the proposed MaxCombo test. Note that this does not allow incorporation of dropouts, so that will be dealt with in simulations.

```
ss <- NULL
for(T in c(18,24,32,40)){
  for(rho in c(0,1)){
    for(gam in c(0,1)){
      n <- SS_HF(H=log(2)/8, HR=c(1,.56), B=32, EPS=6, AT=15, TAU=T-15,
                   P=rho, Q=gam, ALPHA=.025, POWER=.9)
      ss <- bind_rows(ss,
                      tibble::tibble("Study Duration"=T,rho=rho,gamma=gam,
                                     N=n$N,Events=n$Events))
    }
  }
}
kable(ss) %>% kable_styling(bootstrap_options = "striped")
```

Study Duration	rho	gamma	N	Events
18	0	0	3160	1699
18	0	1	1318	709
18	1	0	7496	4028
18	1	1	1864	1002
24	0	0	1094	755
24	0	1	570	394
24	1	0	2650	1829
24	1	1	760	525
32	0	0	628	511
32	0	1	364	296
32	1	0	1712	1392
32	1	1	490	399
40	0	0	496	439
40	0	1	302	268
40	1	0	1502	1329
40	1	1	420	372

We note two things in the above results. First, the G(0,1) always results in the smallest sample size and event count requirement among the four tests considered. Second, we select a study duration of 32 months given the steep increase in sample size for smaller study durations and allowing 2 x control median minimum follow-up.

3 Fixed design sample size

3.1 Computing correlations and significance adjustment

For the 32-month design, we us a large simulation with no treatment effect and the same enrollment, failure rate distribution, dropout rate and follow-up duration to approximate the correlations needed to compute the likely significance-level adjustment to ensure an overall Type I error of 2.5%. We begin by setting up a variety of parameters for the simulations.

```
set.seed(3287)
duration <- 15 # enrollment duration
cutdate <- 32
N <- 50000 # arbitrary sample size
strata<-tibble(Stratum="All",p=1)  # no stratification
block<-c(rep("Control",2),rep("Experimental",2)) # block size of 4
enrollRates<-tibble(rate=N/duration,duration=duration) # constant enrollment
failRates<-tibble(Stratum=rep("All",2), # single stratum
                 period=c(1,1), # single period
                 Treatment=c("Control","Experimental"), # treatments
                 duration=c(1,1), # these will be ignored for this example
                 rate=log(2)/c(8,8)) # hazard rates for control and experimental
dropoutRates=tibble(Stratum=rep("All",2), # constant dropout rates
                    period=c(1,1),
                    Treatment=c("Control","Experimental"),
                    duration=c(1,1),
                    rate=c(.001,.001))
```

Now we perform our large single simulation to approximate correlations needed to adjust the MaxCombo test.

```
# simulate a single trial under null hypothesis of no treatment effect
sim0 <- simPWSurv(n=N,strata=strata,block=block,enrollRates=enrollRates,
                  failRates=failRates,dropoutRates=dropoutRates)

# compute correlation estimate for FH test
result <- sim0 %>%
          cutData(32) %>%
          tensurv(txval="Experimental") %>%
          tenFHcorr(rg=tibble(rho=c(0,0,1,1),gamma=c(0,1,0,1)))
kable(result %>% select(-Z)) %>% kable_styling(bootstrap_options = "striped")
```

rho	gamma	V1	V2	V3	V4
0	0	1.0000000	0.8642665	0.9125453	0.9395603
0	1	0.8642665	1.0000000	0.5829535	0.8920530
1	0	0.9125453	0.5829535	1.0000000	0.7923315
1	1	0.9395603	0.8920530	0.7923315	1.0000000

Next we solve for a nominal Z-value cutoff for the MaxCombo test using the correlation computed above. Rather than build a root finding function here, we tried a few values of `Z` in the following to get an appropriate cutoff to control Type I error. We chose one with a nominal p=0.024 to give some margin for error in simulations used by the GenzBretz algorithm in the `pmvnorm()` function to approximate Type I error.

```
corr <- as.matrix(result %>% select(V1,V2,V3,V4))
Zcutoff <- -2.286
p <- 1-pmvnorm(upper=rep(Inf,4),lower=rep(Zcutoff,4),corr=corr,
               algorithm=GenzBretz(maxpts=50000,abseps=.00001))[1]
p
```

```
## [1] 0.02398832
```

3.2 Sample size derivation

Now we consider the sample size and event count for a 32-month design assuming a delayed effect and an alpha using the adjusted cutoff above. We also assume a slightly larger effect size after the delay to get the sample size and effect count slightly smaller than that we have been studying in our examples to allow some increase when we add an interim analysis. Given that the above does not allow incorporation of the dropout rate, we increase the sample size by 3% from 442 to 456 in order to ensure the targeted number of events accrue in the desired timeframe.

```
n <- SS_HF(H=log(2)/8, HR=c(1,.56), B=32, EPS=6, AT=15, TAU=17,
                    P=0, Q=1, ALPHA=pnorm(Zcutoff), POWER=.9)
n %>% kable() %>% kable_styling()
```

FollowUp	HRpre	HRpost	N	N1	N2	Events	EventsPost
17	1	0.56	442	221	221	360	180

3.3 Simulating Type I error for the fixed design with the MaxCombo test

We verify the above Type I error approximation using simulation.

```
nsim <- 40000
events <- 360
N <- 456
Zvals <- NULL
pMC<-NULL
measures <- NULL
```

```
rho<-c(0,0,1,1)
g<-c(0,1,0,1)
for(i in 1:nsim){
  sim0 <- simPWSurv(n=N,strata=strata,block=block,enrollRates=enrollRates,
                    failRates=failRates,dropoutRates=dropoutRates)
    # cut date at max of event count and 16 month follow-up requirement
    cutdate <- max(getCutDateForCount(sim0,events),max(sim0$enrollTime)+16)
    y0 <- cutData(sim0,cutdate)
    # Compute Cox HR and upper CI
    cox <- survival::coxph(Surv(tte, event) ~ Treatment + strata(Stratum),
          data = y0)
    hr <- exp(cox$coefficients)
    se <- sqrt(cox$var)
    uci <- as.numeric(exp(cox$coef + qnorm(.975) * se))
    # Compute component Fleming-Harrington tests for MaxCombo
    Z <- tenFHcorr(tensurv(y0,txval="Experimental"), rg = tibble(rho=rho,gamma=g))
    # Accumulate Fleming-Harrington simulations
    Zvals <- bind_rows(Zvals, Z %>% mutate(sim=i))
    # Accumulate p for MaxCombo
    p <- 1-pmvnorm(lower = rep(min(Z$Z), nrow(Z)),
                   corr = as.matrix(select(Z, -c(rho, gamma, Z))),
                   algorithm = GenzBretz(maxpts = 25000, abseps=.001))[1]
    # if p-value is small, compute more accurately
    if (p < 0.03) p <-
      1-pmvnorm(lower = rep(min(Z$Z), nrow(Z)),
                corr = as.matrix(select(Z, -c(rho, gamma, Z))),
                algorithm = GenzBretz(maxpts = 50000, abseps=.00001))[1]
    pMC <- bind_rows(pMC,
                   tibble(sim=i,Events=sum(y0$event),Time=cutdate,
                          p=p,
                          HR=hr,
                          uci=uci
                   ))
}
```

The estimated Type I error from this simulation is

```
mean(pMC$p<=.025)
```

```
## [1] 0.0243
```

3.4 Simulating power for the fixed design with the MaxCombo test

We simulate power in the same fashion.

```
nsim <- 1000
Zvals <- NULL
pMC<-NULL
measures <- NULL
rho<-c(0,0,1,1)
g<-c(0,1,0, 1)
for(i in 1:nsim){
  sim0 <- simPWSurv(n=N,
              strata=tibble(Stratum="All",p=1),  # no stratification
              block=c(rep("Control",2),rep("Experimental",2)), # block size of 4
              enrollRates=tibble(rate=N/duration,duration=duration), # constant enrollment
              failRates=tibble(Stratum=rep("All",4),
                               period=c(1,2,1,2),
                               # treatments
                               Treatment=c(rep("Control",2),rep("Experimental",2)),
                               # 6's represent delay period, 1's are ignored
                               duration=c(6,1,6,1),
                               # hazard rates; only last one for experimental is different
                               rate=log(2)/8*c(1,1,1,.56)),
              dropoutRates=tibble(Stratum=rep("All",2), # constant dropout rates
                                  period=c(1,1),
                                  Treatment=c("Control","Experimental"),
                                  duration=c(1,1),
                                  rate=c(.001,.001))
          )
    # cut date at max of event count and 16 month follow-up requirement
    cutdate <- max(getCutDateForCount(sim0,events),max(sim0$enrollTime)+16)
    y0 <- cutData(sim0,cutdate)
    # Compute Cox HR and upper CI
    cox <- survival::coxph(Surv(tte, event) ~ Treatment + strata(Stratum),
          data = y0)
    hr <- exp(cox$coefficients)
    se <- sqrt(cox$var)
    uci <- as.numeric(exp(cox$coef + qnorm(.975) * se))
    # Compute component Fleming-Harrington tests for MaxCombo
    Z <- tenFHcorr(tensurv(y0,txval="Experimental"), rg = tibble(rho=rho,gamma=g))
    # Accumulate Fleming-Harrington simulations
    Zvals <- bind_rows(Zvals, Z %>% mutate(sim=i))
    # Accumulate p for MaxCombo
    pMC <- bind_rows(pMC,
                   tibble(sim=i,Events=sum(y0$event),Time=cutdate,
                          p=1-pmvnorm(lower = rep(min(Z$Z), nrow(Z)),
                                      corr = as.matrix(select(Z, -c(rho, gamma, Z))),
                                      algorithm = GenzBretz(maxpts = 25000,
                                                            abseps = .0005,releps=.01))[1],
                          HR=hr,
                          uci=uci
                   ))
}
```

The estimated power from this simulation is as targeted.

```
mean(pMC$p<=.025)
```

```
## [1] 0.898
```

3.5 Proportional hazards

We examing a proportional hazards sample size with similar power. With 360 events and the adjusted alpha, a logrank with HR=0.687 has approximately 90% power, which will be a slightly conservative estimate of the MaxCombo power; the approximation here uses the Schoenfeld (1981) approximation.

```
nEvents(hr=c(.687,.69,.692),n=360,alpha=pnorm(Zcutoff))
```

```
## [1] 0.8989438 0.8914394 0.8862379
```

4 Group sequential design

4.1 Correlations

Here we add an interim analysis to demonstrate how to compute the correlation between all tests computed in order to properly adjust bounds to control Type I error as desired. We consider only a single interim analysis to simplify the demonstration. We also consider only a logrank test at the interim to not only simplify with a correlation adjustment for fewer tests, but also have a more stringent criterion for early stopping for efficacy compared to a test that depends on weighting.

Suppose we analyze after 75% of 360 events, or 270 events. Here we compute the correlation between a logrank at interim analysis and each of 4 tests: logrank, G(0,1), G(1,0), G(1,1). This will require intermediate calculation of these 4 tests at the interim and the final analysis. This covariance between logrank and the weighted tests at interim is the same as the covariance of the interim logrank with the final weighted tests. We use the last instance of the trial simulation above. The covariance from the final tests above is computed as follows:

```
cutdate <- max(getCutDateForCount(sim0,events),max(sim0$enrollTime)+17)
y0 <- cutData(sim0,cutdate)
FinalCov <- as.matrix(tenFHcorr(tensurv(y0,txval="Experimental"),
                                rg = tibble(rho=rho,gamma=g),corr=FALSE)[,4:7])
FinalCov
```

```
##             V1        V2        V3        V4
## [1,] 87.95367 34.475691 53.477978 16.228911
## [2,] 34.47569 18.246780 16.228911  7.294164
## [3,] 53.47798 16.228911 37.249067  8.934746
## [4,] 16.22891  7.294164  8.934746  3.398717
```

Next, we make a cut for the interim analysis after 270 events.

```
y0  <- cutDataAtCount(sim0, 270)
IACov <- as.matrix(tenFHcorr(tensurv(y0,txval="Experimental"),
                             rg = tibble(rho=rho,gamma=g),corr=FALSE)[,4:7])
IACov
```

```
##             V1        V2        V3        V4
## [1,] 66.94123 20.585218 46.356014 11.809036
## [2,] 20.58522  8.776182 11.809036  4.447009
## [3,] 46.35601 11.809036 34.546978  7.362027
## [4,] 11.80904  4.447009  7.362027  2.444815
```

The covariance between the interim logrank and the final tests is:

```
IACov[1,]
```

```
##       V1       V2       V3       V4
## 66.94123 20.58522 46.35601 11.80904
```

The full covariance matrix for the interim logrank and final logrank, G(0,1), G(1,0) and G(1,1) is thus

```
FullCov <- rbind2(IACov[1,],FinalCov)
FullCov <- cbind2(matrix(c(FullCov[1,1], as.vector(FullCov[1,])), ncol=1),
                  FullCov
                 )
FullCov
```

```
##                      V1        V2        V3        V4
## [1,] 66.94123 66.94123 20.585218 46.356014 11.809036
## [2,] 66.94123 87.95367 34.475691 53.477978 16.228911
## [3,] 20.58522 34.47569 18.246780 16.228911  7.294164
## [4,] 46.35601 53.47798 16.228911 37.249067  8.934746
## [5,] 11.80904 16.22891  7.294164  8.934746  3.398717
```

Converting this to a correlation matrix, we get the covariance for the standardized Z-values of the tests.

```
cov2cor(FullCov)
```

```
##                      V1        V2        V3        V4
## [1,] 1.0000000 0.8724085 0.5890003 0.9283282 0.7829069
## [2,] 0.8724085 1.0000000 0.8605832 0.9343085 0.9386527
## [3,] 0.5890003 0.8605832 1.0000000 0.6224989 0.9262430
## [4,] 0.9283282 0.9343085 0.6224989 1.0000000 0.7940851
## [5,] 0.7829069 0.9386527 0.9262430 0.7940851 1.0000000
```

4.2 Interim spending

The interim spend and corresponding Z-value cutoff is

```
spend <- sfLDOF(alpha=.025,t=270/360,param=NULL)$spend
spend
```

```
## [1] 0.009649325
```

```
qnorm(spend)
```

```
## [1] -2.339711
```

4.3 Testing bounds

Now we create a function to search for a cutoff for the final analysis that preserves 0.025 total Type I error spend for the interim and final. We note that accuracy for the `pmvnorm` function must be set to provide sufficiently accurate results for this computation. We use a maximum number of interations for the root-finding routine and then test our result for accuracy.

```
corrmat <- cov2cor(FullCov)
errval <- function(x, corr=corr, z1,alpha=.025){
  1-pmvnorm(lower=c(qnorm(spend),rep(x,4)),upper=rep(Inf,5),mean=rep(0,5),corr=corrmat,
            algorithm=GenzBretz(maxpts=50000,abseps=.00001))[[1]]
}
errval(-2.305, z1=qnorm(spend))
```

```
## [1] 0.02467049
```

```
pnorm(-2.305)
```

```
## [1] 0.01058329
```

```
Zcutoff2 <- -2.305
```

Thus, the final cutoff is a Z-value of -2.305 which has a nominal standard normal p-value of 0.0105833.

5 Conclusions

While not as simple as deriving a design for a logrank test under proportional hazards, the methods presented hear provide a practical way to design a group sequential trial with a MaxCombo test as well as simulation tools that can be used to verify design properties. This can result in considerable sample size savings or increase in power for many cancer trials where there is a great unmet medical need and limited patients to enroll.

6 Session information

```
sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: i386-w64-mingw32/i386 (32-bit)
## Running under: Windows 10 x64 (build 17763)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] mvtnorm_1.1-0       kableExtra_1.1.0    dplyr_0.8.5
## [4] survival_3.1-8      simtrial_0.1.7.9005 gsDesign_3.1.1
## [7] ggplot2_3.3.0
##
## loaded via a namespace (and not attached):
##  [1] withr_2.2.0       readr_1.3.1       rvest_0.3.5       tidyselect_1.1.0
##  [5] lattice_0.20-38   pkgconfig_2.0.3   xml2_1.3.2        compiler_3.6.3
##  [9] stringr_1.4.0     viridisLite_0.3.0 xtable_1.8-4      labeling_0.3
## [13] Rcpp_1.0.4.6      httr_1.4.1        plyr_1.8.6        tools_3.6.3
## [17] rmarkdown_2.1     R6_2.4.1          purrr_0.3.4       knitr_1.28
## [21] scales_1.1.1      assertthat_0.2.1  digest_0.6.25     gtable_0.3.0
## [25] evaluate_0.14     Matrix_1.2-18     stringi_1.4.6     rstudioapi_0.11
## [29] farver_2.0.3      htmltools_0.4.0   hms_0.5.3         munsell_0.5.0
## [33] grid_3.6.3        lifecycle_0.2.0   colorspace_1.4-1  glue_1.4.0
## [37] rlang_0.4.6       magrittr_1.5      ellipsis_0.3.0    splines_3.6.3
## [41] vctrs_0.3.0       yaml_2.2.1        testthat_2.3.2    crayon_1.3.4
## [45] xfun_0.14         tidyr_1.0.3       pillar_1.4.4      webshot_0.5.2
## [49] tibble_3.0.1
```

References

Hasegawa, Takahiro. 2014. “Sample Size Determination for the Weighted Log-Rank Test with the Fleming–Harrington Class of Weights in Cancer Vaccine Studies.” *Pharmaceutical Statistics* 13 (2): 128–35.

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

Schoenfeld, David. 1981. “The Asymptotic Properties of Nonparametric Tests for Comparing Survival Distributions.” *Biometrika* 68 (1): 316–19.