CONSIDERATIONS FOR DESIGN AND ANALYSIS OF TRIALS WITH POSSIBLY NON-PROPORTIONAL HAZARDS

October 16, 2018
Keaven M. Anderson, Ph.D. and Satrajit Roychoudhury
Merck Research Laboratories and Pfizer
Disclaimer

While the authors are members of the Non-Proportional Hazards (NPH) Working Group, any mistakes and opinions should be considered those of the authors. Also, this work does not represent a company position for either Merck or Pfizer.
Acknowledgements

• Members of the Cross Pharma NPH Working Group

  • Leadership Team
    – CoLeaders: Renee Iacona (AZ), Tai-Tsang Chen (BMS)
    – Design and Analysis Workstream: Keaven Anderson (Merck), Satrajit Roychoudhury (Pfizer), Tianle Hu (Lilly), Ray Lin (Roche)
    – Endpoint Workstream: Jane Qian (Abbvie), Dominik Heinzmann (Roche)
    – Simulation Team: Julie Cong (B&I), Tianle Hu (Lilly)
    – Case Studies: Pralay Mukhopadhyay (AZ)

• Organizations represented in the Working Group
  – AZ, BMS, Merck, Boehringer Ingelheim, Novartis, Lilly, Abbvie, Genentech, Roche, Bayer, Janssen, Takeda, Amgen, Pfizer, GSK, Celgene, Sanofi, Johnson & Johnson, and FDA
Outline

• Background
• Methods studied
• Simulation summary
• Recommendations for testing and estimation
• Design considerations
• When do the results break down?
• Summary
BACKGROUND

Based on slides from Rajeshwari Shridhara, FDA

https://healthpolicy.duke.edu/events/public-workshop-oncology-clinical-trials-presence-non-proportional-hazards
Time to Event Analysis in Randomized Clinical Trials

• Endpoint Examples: Overall survival (OS), Progression-free Survival (PFS), Recurrence-free Survival (RFS), Disease-free Survival (DFS), etc.

• Most common approach in design of clinical trials with time to event endpoint:
  – Fix chance of false positive conclusion (alpha)
  – Fix chance of winning or detecting benefit if it exists (power of the test)
  – Define what treatment effect is meaningful (alternative hypothesis to null hypothesis of no effect)
  – Assume relative treatment effect (hazard ratio) is constant over time
Standard Time to Event Analysis

Assuming constant relative treatment effect over time,

• Comparison of survival curves using Log-rank test (Non-parametric test)
  – Estimated median survival provides a summary of the survival curve (i.e., on an average, 50% of events observed before the median time)

• Test hypothesis and estimate relative treatment effect using Cox-proportional hazards model
  – Hazard ratio provides an average relative effect over time

• Power to test the hypothesis reduces as relative effect changes over time (violation of the constant effect assumption)
Kaplan-Meier Curves of Progression-free Survival Based on IRAC Assessment (ITT Population) Between Arms Rd Continuous, Rd18 and MPT (Lenalidomide product label)
Nivolumab 2nd line non squamous mNSCLC: PFS analysis
Comparing Treatments in the Presence of Crossing Survival Curves: An Application to Bone Marrow Transplantation

Kaplan-Meier estimate of DFS for Follicular Lymphoma by transplant source
Pemetrexed in Mesothelioma (product label)

Median Survival = 12.1 mos
Hazard Ratio = 0.77
Log rank p-value = 0.020
Percent censored = 32

Median Survival = 9.3 mos
P-values: Max-combo test vs. log-rank test

- The use of weights in the max-combo test suggests that some events are “more important” than others. How to justify it?
- Source: Lijun Zhangj, FDA Duke-Margolis slide set
Challenges

• When the assumption of constant HR is not true,
  – Cox-proportional hazard model is inappropriate
  – KM estimate of median survival may not be an optimal measure to summarize the results

• What is an optimal analysis method to test treatment benefit and how can we summarize the benefit?
  – Many methods have been proposed in literature; each have advantages and limitations
  – Multiple approaches may be necessary to summarize results
FDA Initiated Collaboration

• FDA recognized the need for collaboration

• Initiated dialogue with the Industry statisticians

• Met in 2016 and subsequently in 2017
  – Concluded that a methodical evaluation of available methods is needed
  – Goal: identify appropriate analysis method for the different patterns of non-proportionality
  – All industries to work together as a team (non-product specific)
  – FDA to participate in this effort
Why are we here today?

• Current practice of using log-rank test and Cox-proportional hazards model not appropriate when relative treatment effect varies over time (Non-proportional Hazards)

• Reasons for observed changes in treatment effect over time may be different in different clinical trials

• What is the best way to evaluate treatment effect?

• What is the best way to summarize an observed treatment effect?

• Working group will be presenting what has been accomplished so far
ANALYSIS METHODS
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 (LRT): detect difference in treatment effect
  – Cox regression (CR): summarize the treatment effect

• Log-rank p-value, hazard ratio, 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?
Log-rank Test and Cox Regression: Fits to All?

- **LRT**: introduced by Nathan Mantel in 1966
- **CR**: introduced by Sir David R. Cox in 1972
- LRT and CR are **closely related**
- LRT is fully nonparametric
  - **asymptotically efficient** for proportional hazards (PH)
  - **substantial power loss** if PH assumption does not hold
- Key assumption for CR: **constant** effect over time
  - treatment effect summarized by hazard ratio (HR)
  - problematic if PH assumption is violated
Analysis and Design Trial with NPH: Key Challenges

• NPH has been discussed extensively in literature
  – alternative methods for hypothesis testing and estimation

• However, application in real life is still rare

• **Main challenge**: NPH type cannot be pre-identified
  – treatment effect profile is unknown at design stage

• **Key questions** for today’s forum: in presence of NPH
  – how to plan primary analysis appropriately?
  – how to design a trial?
  – how to efficiently communicate the results with non-statisticians?
Choice of Primary Analysis in Confirmatory Trials

- 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)
Choice of Primary Analysis Methods

Choice of Methods

- **Rank based Test**
  - LRT
  - Weighted LRT

- **Combination Test**
  - Combination of weighted LRT
  - Breslow test

- **KM based Test**
  - Weighted KM test
  - Restricted mean survival time (RMST)
Weighted Log-rank Test

Fleming and Harrington proposed a class of weighted log-rank test (FH) based on the $G^{\rho,\gamma}$ family

Assign weight to events

$$W_n(t) = (S_n(t))^{\rho}(1 - S_n(t))^\gamma$$

Values of $\rho$ and $\gamma$ implies

- $\rho > 0$, $\gamma = 0$: early difference
- $\rho = 0$, $\gamma > 0$: late difference
- $\rho > 0$, $\gamma > 0$: mid difference
- $\rho = 0$, $\gamma = 0$: log-rank test
Combination Test

• Major difficulty for FH LRT:
  – specification of $\rho$ and $\gamma$ parameter: mis-specification may imply a loss of power

• Possible alternative: **Combination test**
  – handles simultaneously a range of NPH types
  – choose the appropriate weight in “adaptive” fashion

• Similar concepts are explored by
  – Yang and Prentice 2010: *Adaptively Weighted log-rank Test*
  – Karrison 2016: *Versatile tests*
  – Garès et. al. 2017: maximal statistics over FH(0,$\gamma$)
Combination of FH Log-rank Test (Max-Combo)

• We have considered two combinations
  – combination of $G^{0,0}$ and $G^{0,1}$ : **Combo 1**
  – combination of $G^{0,0}$, $G^{0,1}$, $G^{1,1}$, $G^{1,0}$ : **Combo 2**

• **Max-Combo test** : largest of the absolute value of the test statistics

• “Adaptive” procedure involving selection of best test statistics: 
  requires multiplicity correction
  – Bonferroni-Holm adjustment (conservative)
  – adjustment using the joint asymptotic distribution of the FH log-rank test statistics (recommended)

• Can be pre-specified easily at protocol stage : satisfies ICH E9 condition
Kaplan-Meier Based Tests

KM plot is a well understood tool showing fraction of patients living for a certain amount of time after treatment.

- Area under KM represents RMST
- The life expectancy of patient over the next 20 months
- Treatment effect (Difference scale) at month 20
- KM based test are based on the difference/ratio between two KM curves

Data cutoff
- Take maximum follow-up in each treatment group
- Minimum of these maxima is cutoff
- Recent justification for this for RMST submitted for publication
Kaplan-Meier Based Tests

• **Weighted Kaplan-Meier test:** (Pepe and Fleming, 1989, 1991)
  - weighted difference of area under KM curves up to a specified cut-off
  - weights are based on KM estimate of censoring
  - need to specify the cut-off: can be affected by censoring

• **Restricted mean survival time (RMST)** (Uno et al 2014)
  - area under the KM plot prior to specific time-point: can be easily interpreted as “life expectancy”
  - treatment effect: difference or ratio of RMST
  - need to specify the cut-off: can be affected by censoring
Other Methods

• **Piecewise log-rank test (Xu. *et al* 2016)**
  – piecewise weighted log-rank test within specified time intervals
  – optimal when weights for earlier events are zero
  – *power/type-I error* greatly affected *if intervals are incorrectly specified*

• **Other combination tests**:
  – **Breslow *et. al. 1984***: combination of log-rank test and test of acceleration
  – **Logan 2008**: combination of log-rank test and milestone survival, it suffers similar problem as other KM based tests

• **Net chance of longer survival**: Buyse (2010), Peron et al (2018)
  – Generalized pairwise comparison
  – Can specify ‘clinically significant’ difference for pairwise evaluations
Reporting Treatment Effect

• When NPH is present: HR depends on time
  – HR or average HR as a single number is less useful
  – what statistics to be reported to quantify treatment effect?
  – how to appropriately pre-specify to meet ICH E9?

• A sequential approach (Royston and Parmer 2010)
  – First step: perform Max-combo test to conclude about the “Null” hypothesis (no treatment effect)
  – Second step: regardless of results in step 1, gather evidence of NPH, possible options
    ▪ Grambsch–Therneau test for PH
    ▪ other graphic diagnostics for confirming PH
  – Third step: choose treatment effect summary based on step 2- treatment effect estimate beyond test statistics

• Net chance of longer survival
  – Interesting with pre-specified cutoff or as a function of minimum important difference?
Choice of Treatment Effect Summary

• If PH assumption is reasonable
  – **HR from Cox regression (CR)** and corresponding 95% confidence interval (CI)
  – secondary analysis: average HR from weighted CR and 95% confidence interval (weight chosen by Max-combo)

• If there is evidence of NPH, the possible metrics
  – **ordinary** and **average HR (Max-Combo)** with 95% CI
  – difference in RMST at max cutoff
  – difference in milestone survival at $t^*$: gain in chance of survival at clinically relevant time point $t^*$ (pre-specified)
  – secondary analysis: piecewise HR and/or piecewise failure rates with 95% CI
SIMULATION STUDIES
Simulation scenarios studied

Simulation results

- MaxCombo had competitive power for all scenarios
- Type I error controlled when survival is equal
- Individual tests performed poorly in at least some scenarios
What is null hypothesis space for weighted logrank?

For weighted logrank, benefit measured as a function of relative failure rates.

- This may not correspond to a survival benefit.
- For increasing weights, this can be out of null hypothesis space.

HR = 1.556 month 0-6

HR = 0.869 after 6 months
Type I error controlled by MaxCombo?

- Underlying survival distribution
  - Controls exponential with median of 15 months ($\lambda=0.046$)
  - Experimental group is piecewise exponential
    - HR=1.556 for 6 months
    - HR=0.869 thereafter
    - Survival curves cross at 30 months
- Enrollment: N=200
  - Constant enrollment rate for 12 months
- Data cutoff: 30 months
- Type I error (1-sided; 10k simulations)
  - MaxCombo: 1.5%
  - MaxCombo also requiring upper CI for HR < 1.1: 0.78%
  - Inflated for FH(0,1): 2.7% (within simulation error)
  - There are potential issues here in some cases
STUDY DESIGN
Design issues

• Trials results often differ from design assumptions

• Results may differ by
  • Degree of effect
  • Delayed timing of effect
    • Delayed separation of survival curves
  • Different effects in unanticipated subpopulations
    • This can result in crossing hazards
  • Diminishing effect over time
    • Converging hazards – maybe of LESS interest here

• How do we design a trial to be powerful across MANY alternatives?
Design philosophy

• Power trial for multiple scenarios
• Find worst-case scenario, e.g.,
  • Minimum effect size of interest (PH)
• Delayed effect
• Early crossing hazards
• Simple approximation of alternatives
  • Piecewise exponential failure
  • Single change point
• No single estimand/estimate is adequate
  • Inconsistent with ICH E9 (R2) estimand recommendations?
Design implementation

• Ensure adequate follow-up
• Robust testing method
• If using MaxCombo
  • Karrison (2016) provides correlations needed to adjust for multiple tests
  • Power for multiple scenarios & select worst-case sample size
    • Use adjusted significance level for components of MaxCombo
      • Modification of Hasegawa (2016) for calculation
      • Power for best MaxCombo component will be conservative
Design: interim analysis (IA) considerations

- Recommend logrank for interim stopping
- Improve regulatory acceptance?
- May wish to use MaxCombo for sensitivity analysis

- Lack of efficacy
  - Are early tests of excess mortality required?
  - Early safety bounds rather than futility bounds
  - Conditional power-based futility: Freidlin and Korn (201?)

- Efficacy testing
  - Delayed effect may result in fast event accumulation
  - Set timing based on events AND follow-up to ensure power
BREAKDOWN AND ESTIMATION EXAMPLES
Breakdown examples

- **MaxCombo fails**
- **Logrank and other tests succeed**

Only MaxCombo succeeds

- **MaxCombo succeeds then fails since upper CI HR > 1.1**

**Proportional hazards example**

**Moderate crossing hazards example**

**Delayed effect example**

**Severe crossing hazards example**
Net chance of longer survival example

- Preferred cutoff may be patient-dependent
- Power not well-studied in our simulations
  - For examples, was not positive other than for PH
- Is this helpful beyond Kaplan-Meier curve?
### Summarizing benefit

**Moderate crossing hazards example**

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Experimental</th>
<th>Control</th>
<th>Estimate (95% CI)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median/HR/logrank</td>
<td>5.594</td>
<td>7.303</td>
<td>0.878 (0.708,1.089)</td>
<td>0.118</td>
</tr>
<tr>
<td>Weighted HR/MaxCombo</td>
<td>NA</td>
<td>NA</td>
<td>0.689 (0.515,0.923)</td>
<td>0.004</td>
</tr>
<tr>
<td>RMST</td>
<td>10.544</td>
<td>9.503</td>
<td>1.041 (-0.767,2.849)</td>
<td>0.130</td>
</tr>
<tr>
<td>RMTL</td>
<td>16.941</td>
<td>17.982</td>
<td>0.942 (0.849,1.046)</td>
<td>0.131</td>
</tr>
<tr>
<td>% favorable by 6 mos</td>
<td>25.244</td>
<td>26.297</td>
<td>-1.054 (-10.415,8.056)</td>
<td>0.593</td>
</tr>
<tr>
<td>Weighted KM</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
<td>0.367</td>
</tr>
</tbody>
</table>
Summarizing benefit: Milestone survival
Moderate crossing hazards example

<table>
<thead>
<tr>
<th>Month</th>
<th>Experimental</th>
<th>Control</th>
<th>Difference (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>66%</td>
<td>71.5%</td>
<td>-5.5% (-14.2%,3.3%)</td>
</tr>
<tr>
<td>6</td>
<td>48.4%</td>
<td>55.2%</td>
<td>-6.8% (-16.2%,2.6%)</td>
</tr>
<tr>
<td>12</td>
<td>34.8%</td>
<td>32.8%</td>
<td>2% (-7%,11%)</td>
</tr>
<tr>
<td>18</td>
<td>27.9%</td>
<td>16.1%</td>
<td>11.8% (3.6%,20.1%)</td>
</tr>
<tr>
<td>24</td>
<td>20.6%</td>
<td>9.4%</td>
<td>11.2% (2.1%,20.3%)</td>
</tr>
</tbody>
</table>
Summarizing benefit: Piecewise exponential failure rates
Moderate crossing hazards example

<table>
<thead>
<tr>
<th>Period</th>
<th>Experimental</th>
<th>Control</th>
<th>HR (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-3 months</td>
<td>0.139</td>
<td>0.113</td>
<td>1.237 (0.88, 1.737)</td>
</tr>
<tr>
<td>3-6 months</td>
<td>0.103</td>
<td>0.086</td>
<td>1.194 (0.754, 1.89)</td>
</tr>
<tr>
<td>6-12 months</td>
<td>0.058</td>
<td>0.087</td>
<td>0.666 (0.419, 1.057)</td>
</tr>
<tr>
<td>&gt;12 months</td>
<td>0.038</td>
<td>0.109</td>
<td>0.350 (0.199, 0.616)</td>
</tr>
</tbody>
</table>
SUMMARY
Potential concerns for alternative methods for regulatory approval

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

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

• MaxCombo useful for non-proportional hazards in metastatic setting
• Important benefit could be missed with other methods
• Proposals are ready for alternatives to logrank/Cox/median
• Sponsors encouraged to submit as supportive
• Further discussion needed to move approaches to primary
References


THANK YOU