Non-Proportional Hazards – So What?

Andrew Stone
BBS spring seminar April 28th 2016
Disclaimer

Andrew Stone is an employee of AstraZeneca LP. The views and opinions expressed herein are my own and cannot and should not necessarily be construed to represent those of AstraZeneca or its affiliates.
Acknowledgements

- Wenmei Huang
- Monica Huhn
- Mary Jenner
- Paul Metcalfe
- Pralay Mukhopadhyay
- Fredrik Ohrn
- Luping Zhao
- (All AZ – Oncology TA or Advanced Analytics Centre)
Justifiably huge excitement about a new class of agents

- As predicted by biology, one of the earliest results suggested a delayed effect
- This raised fundamental questions about the design and analysis of data from this class, especially for situations where the treatment effect was not so large

One fundamental question: is the hazard ratio (HR) interpretable in the presence of non-proportional hazards (NPH)?

- Influential publication*

‘When the PH assumption is violated (i.e., the true hazard ratio is changing over time), the parameter actually being estimated by the Cox procedure may not be a meaningful measure of the between group difference; it is not, for example, simply an average of the true hazard ratio over time.’

Really?? Let’s examine this assertion.

* Uno H, J Clin Oncology 2014 2380-5
Is this any different to multicentre clinical trials?

• Quantitative treatment-by-centre interaction
  • We’re quite happy to describe the treatment benefit as the average over centres even if there is statistical evidence that the benefit differs across centres

• NPH = quantitative treatment-by-trial interaction
  • Acknowledge, but describe overall benefit as the average treatment effect over time
  • Likely with a large enough trial there would always be evidence of NPH

• But how do we estimate the average HR?
The HR estimated from a standard cox/log-rank is the average HR – with all events weighted equally

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Hazard rate</th>
<th>HR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>New</td>
<td>Control</td>
</tr>
<tr>
<td>0-3</td>
<td>0.1</td>
<td>0.2</td>
</tr>
<tr>
<td>3-6</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>HR estimated by Cox</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>Weighted average of piecewise HR</td>
<td>0.86</td>
<td></td>
</tr>
</tbody>
</table>

• HR = geometric mean of piecewise HRs, weighted proportional to no. of events per period

\[
\overline{HR} = \exp \left( p_1 \ln(HR_1) + p_2 \ln(HR_2) \right)
\]

• where \( p_1 \) and \( p_2 \) are proportion of events per period

• Noting that all patients, with events, are treated as equally important in terms of increasing life
ln(HR) ~ U/V*,

- where $U = \sum (d_{ij} - \frac{n_{ij}d_j}{n_j})$ the usual log-rank denominator
- and $V = \sum \frac{n_{ij}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j-1)} \sim e/4$ the usual log-rank numerator which is equal to the reciprocal of the variance for the ln(HR) with e the total of events

U and V can be partitioned into summations before and after a change in HR and noting that the above implies U~ (e/4) . ln(HR)

Therefore the overall lnHR

\[
\frac{(U_1+U_2)}{(V_1+V_2)} = \frac{(e_1/4 \cdot \ln(HR_1) + e_2/4 \cdot \ln(HR_2))}{(e_1/4 + e_2/4)} = p_1 \ln(HR_1) + p2 \ln(HR_2)
\]

Ah but...

- Cox ‘assumes’ proportional hazards
  - Assumes an unfortunate word as implies, with lack of PH, the test is somehow not valid
  - ‘Assume’ actually means ‘most powerful’ when the alternative is NPH
  - Under H0 by definition we have PH anyway

- Incidentally, many times I’ve heard it stated, incorrectly, the log-rank (LR) doesn’t assume PH as makes less ‘assumptions’ as it’s fully non-parametric
  - Cox and LR will always give results very close
  - Can be made to be identical if both based on a score test, use same method to handle ties and stratified by the same factors

- But should acknowledge in interpretation that with further follow-up that treatment effect will change
  - Importance of follow-up
  - This applies equally to the alternatives proposed
However, regardless of proportionality ……

Same meaning clinically?
We need to supplement with absolute benefit

\[ \text{Exponential distn} \]
Control median = 5

\[ \text{Weibull distn} \]
\[ S(t) = \exp(-0.028t^2) \]
Control median = 5

\[ \text{HR} = 0.5 \]
\[ \Delta \text{Median} = 5 \]
\[ \Delta \text{Mean} = 7.2 \]

\[ \text{HR} = 0.5 \]
\[ \Delta \text{Median} = 2.1 \]
\[ \Delta \text{Mean} = 2.2 \]

\[ \text{NOTE: A kaplan-meier of the ranks looks identical for both distributions, as HR based on relative ranking not actual times} \]
HR remains meaningful and the primary measure of effect

But supplemental measures needed

But what?
Medians normally used for absolute benefit, yet we know they’re a lousy measure

• Medians
  X What happens afterwards has no bearing
  X For PFS in particular: random steps on KM curve and dependence on timing of scans for PFS

• So what are the alternatives?
‘Landmark analyses’ – easy to understand but...

- Could compare the proportion of patients surviving 1 year
  - Timepoint pre-specified
  - Need to use KM estimate and adjust variance accordingly\(^1\)
  - Note ratio of \(\ln(S(t)) =\) ratio of average hazards, different to average hazard ratio

- Very easy for clinicians and patients to understand
  - Your chance of surviving for a year is increased by \(x\%\)

- Could be more powerful than HR with NPH
  - Depends on separation and no. of events after the timepoint
  - Note if \(~\)PH then always less powerful than HR as less events included\(^2\)

What about mean survival

• Easy to understand
  • Standardly used with continuous data
  • Same as AUC of KM curve

• But....
  • Requires a high proportion of events (ie high maturity and little censoring)
  • Could be unduly influenced by a few events

• As a requirement therefore would delay access to medicines

Figure 1. Kaplan-Meier Plot of Progression-Free Survival Time as Determined by the IRC

Precedent for using in labelling Panitumumab

‘The mean PFS was 96 days in the Vectibix arm and 60 days in the BSC-alone arm.’
Restricted Mean gaining popularity

• Suggested by authors at the Medical Research Council, UK¹
• Idea to restrict inference to period with PH
• Calculate mean during that period, adjusting for censoring
• Hard concept to convey
• If you progress within 2 years you’ll progress 3 months later on average

‘Some evidence of non-proportional hazards was noted (p=0.06) and the restricted mean survival time over 2 years was 12.5 months (11.7–13.4) in arm C and 9.4 months (8.6–10.2) in arm A.’²

Figure 2: Kaplan-Meier plot of progression-free survival over 2 years. Vertical reference line shows the median time to completion of the chemotherapy phase. Number at risk every 6 months shown with the number of failure events in parentheses, after the time in which the number at risk was calculated.

¹ Royston and Parmar BMC Medical Research Methodology 2013, 13:152
² Ledermann et al.. Lancet Vol 387 March 12, 2016
Mean restricted regardless of period of proportionality

Progression-Free Survival

<table>
<thead>
<tr>
<th></th>
<th>Nivolumab</th>
<th>Docetaxel</th>
</tr>
</thead>
<tbody>
<tr>
<td>nPFS, mo</td>
<td>1.5</td>
<td>2.8</td>
</tr>
<tr>
<td>(95% CI)</td>
<td>(2.1, 4.9)</td>
<td>(2.1, 3.5)</td>
</tr>
<tr>
<td>HR</td>
<td>0.62</td>
<td>0.61</td>
</tr>
<tr>
<td>(95% CI)</td>
<td>(0.47, 0.81)</td>
<td>(0.47, 0.81)</td>
</tr>
<tr>
<td>p</td>
<td>0.0004</td>
<td>0.0004</td>
</tr>
</tbody>
</table>

Estimated from curves:

- RMST difference with truncation point at 12 months ~ 1.5
- RMST difference with truncation point at 18 months ~ 2.4

Presented By David Spigel at 2015 ASCO Annual Meeting; RMST calculated by digitization of curves
Properties of RMST – subject of study at AZ

HR and RMST similar power, expected as HR ~ relative AUC of KM of ranks?

$X\text{-axis} = T/\text{median}$ where $HR=1 < T$ with $T=\text{time to separation of curves}$

Simulations by Luping Zhao
Can we make better use of parametric approaches?

With PH
• The Event Time Ratio (ETR\(^1\)) could be estimated from a weibull accelerated failure time model
  • where each arm has the same shape parameter and thus would also have proportional hazards
• For all percentiles, the treatment effect is delayed by a common % = ETR

Regardless of proportionality, mean survival times can be expressed as a function of parameters specific to each treatment arm
• See Ellis\(^2\) for means and variances
• Advantage: these would represent overall predicted means rather than those restricted to a timeperiod. To be studied further

\(^1\) Carroll KJ Controlled Clinical Trials 24 (2003) 682–701
\(^2\) Ellis S Contemporary Clinical Trials 29 (2008) 456–465
All of this very important: ASCO & ESMO Value Framework:
Scenario: Advanced Disease with OS as Primary

### The ASCO Value Framework: Advanced Disease

**Step 1:** Determine the regimen’s CLINICAL BENEFIT

<table>
<thead>
<tr>
<th>Overall Survival (OS) reported?</th>
<th>OS Score</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes, Assign an OS Score (1 through 5 as shown below) and multiply by 16. Write this number in the box labeled, “OS Score.” Proceed to 1.D.</td>
<td>0%–24%</td>
<td>25%–49%</td>
<td>50%–75%</td>
<td>76%–100%</td>
<td>At double the median OS of new regimen, there is a 50% improvement in the fraction of patients surviving</td>
<td></td>
</tr>
<tr>
<td>Improvement in median OS (% change in median OS)</td>
<td>NO. Proceed to 1.B.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

**Form 2a:** for therapies that are not likely to be curative with primary endpoint of OS

**IF median OS with the standard treatment is ≤ 1 year**

- **Grade 4**
  - HR ≤ 0.65 AND Gain ≥ 3 months
  - Increase in 2 year survival alone ≥ 10%

- **Grade 3**
  - HR ≤ 0.65 AND Gain 2.5–2.9 months
  - Increase in 2 year survival alone 5–<10%

- **Grade 2**
  - HR > 0.65–0.70 OR Gain 1.5–2.4 months
  - Increase in 2 year survival alone 3–<5%

- **Grade 1**
  - HR > 0.70 OR Gain <1.5 months
  - Increase in 2 year survival alone <3%

**IF median OS with the standard treatment > 1 year**

- **Grade 4**
  - HR ≤ 0.70 AND Gain ≥ 5 months
  - Increase in 3 year survival alone ≥ 10%

- **Grade 3**
  - HR ≤ 0.70 AND Gain 3–4.9 months
  - Increase in 3 year survival alone 5–<10%

- **Grade 2**
  - HR > 0.70–0.75 OR Gain 1.5–2.9 months
  - Increase in 3 year survival alone 3–<5%

- **Grade 1**
  - HR > 0.75 OR Gain <1.5 months
  - Increase in 3 year survival alone <3%

Both frameworks also includes grading for toxicity and/or QOL and cost. This presentation focuses on efficacy grading only.
Impact on Trial Design
Sizing with a delayed treatment effect – for future reference

Assume
\[ \text{HR}_1 = 1 \quad t < T, \quad \text{HR}_2 = x < 1 \quad t \geq T, \]
where T denotes the lag-time until there is a benefit of therapy and \( \text{HR}_2 \) the hazard ratio (experimental : control) before and after the lag respectively.

The overall average HR is given by \(^1,^2\):
\[ \overline{HR} = \exp \left( p_2 \ln(\text{HR}_2) \right) \]
Where \( p_2 \) is the proportion of events observed before and after the lag-time respectively.
Therefore power will increase as \( p_2 \) increases

Then assume patients are recruited according to \(^3\): \[ G(s) = \frac{s^k}{B^k} \]
k=2 often approximates reality well
For a given follow-up:
\[ p(\text{event by time } t) = \left( \frac{\min(t, B)}{B} \right)^k - \frac{k}{B^k} \int_0^{\min(t,B)} s^{k-1} S(t-s) \, ds \]
\( \overline{HR} \) can then be calculated and together with \( n \), the total no. of events, the following can be re-arranged to estimate power
\[ n = \frac{(1 + r)^2 \cdot \left[ \Phi^{-1}(1 - \alpha/2) + \Phi^{-1}(1 - \beta) \right]^2}{r \cdot \ln^2(\overline{HR})} \]
Where \( r = \) randomisation ratio (eg \( r=1 \) with 1:1 randomisation)

\(^3\) Carroll KJ (2009), Pharmaceutical Statistics, 8, 333–345.  A closed form solution is presented with \( T=0 \), exponential and integer k
Example survival curves with $T=0$ and 2

HR=1 before separation

HR=0.5 without delayed separation

HR=0.5 after separation
Adverse impact on power if delay is not accounted for

1:1 randomization; assumes e/0.71 patients are recruited where e = no. of events
Fixed 15 months accrual time (uniform);
Median OS (control)=7 months; 2-sided type I error =0.05;

Simulations by Wenmei Huang
Unlike PH, power dependent on maturity

With NPH power increases as proportion of events that occur after time-lag increases

PH - H1: HR=0.6
NPH – H1: HR=1 for t≤4, HR=0.4 for t>4
In this case, power coincides when trials have ~50% maturity e.g. 320 patients
An alternative approach to futility analyses maybe needed

• For example, final analysis to be conducted with 194 events out of 274 patients (71% maturity) ¹

• If the futility analysis ² is planned after 97 events, then either this analysis could be performed
  a) After the first 97 events occur
  b) Alternatively, only including, and after the first 97 events have occurred amongst the first 137 patients recruited (71% maturity same as final analysis)

• If $T$, the lag-time =2, then the probability of false negative is
  • 11% for option a)
  • 5% for option b)

¹ Median OS in the control arm of 7 months, 1:1 randomization; uniform accrual of 30 patients per month; target HR of 0.625; $T=0$.
² Total events adjusted to 194 events with LanDeMets OBF beta, 10%, spending. Futility if interim HR $> 0.948$. 
Log-rank test – can we do better?

- The log-rank test weights each event equally
- There exist alternatives with different weight per event
- One alternative is to use the $G^{\rho,\tau}$ class$^1$ of weighted log-rank tests
- Where:
  - $\rho=0$, $\tau=0$ corresponds to the log-rank
  - $\rho=0$, $\tau=1$, weights proportionately to (1-$S(t)$), estimated from KM, hence more weight to later events

5000 simulations with: 193 events from 266 patients, HR follows (1) with $x = 0.625$ for different $T (=0, 2, 4$ and 6-month, respectively), 15 month accrual, median OS=7, 2-sided $\alpha=5\%$
But should we??

• The use of unequal weights implies increasing survival is more important for some patients than others

• That’s OK if we can identify those patients before they’re dosed

• But if not, why would it be more important to increase survival of those patients who have the better prognosis?

• However, if there was evidence of a cure:
  • If no evidence of harm to any patients, the overall population may have a +ve B/R if no overall average effect but an important proportion of patients were cured
How well do cure rate models work?

• Parametric mixture models investigated

\[ P(T_{event} > t) = p + (1 - p)e^{-(t/T)^b} \]

\( p = \) proportion cured and weibull survival in non-cured

• Long term cured model derived based on nivolumab melanoma data

• Two questions investigated
  1) How do you know whether you’ve correctly identified the cure rate?
  2) How much follow-up do you need?
Models will always provide an estimated cure fraction. But not necessarily the correct one!

- Cure fraction badly over-estimates truth if analysed too early
  - 95% CI excludes truth (not shown)
- To the extent that lower CI for cure fraction excludes the truth

True curve simulated from

Simulations by Monika Huhn & Paul Metcalfe
How long should we follow patients until we can be confident of estimated cure rate?

• First of all would require a cause-specific survival analysis
  • Censoring non-cancer deaths
• One possibility, proportion of uncensored observations with an event in the interval \([t^* - (t-t^*), t^*]\), where \(t^*\) = latest event (uncensored) time, \(t\) = largest time (event or censored).

An aside: \(q\) uses a denominator of the total no. of observations. Whereas if the denominator was the number of uncensored observations would have better properties
• max value =1 independent of cure rate
Just finally – what’s the emerging picture in terms of presence of delayed effect?
Careful how PFS curves are (mis?) interpreted

• Care required when examining presence of time-lag with PFS data

• Have seen KM curves misinterpreted a few times

• In the example opposite:
  • At the first scheduled scan, 9 weeks, there is a clear difference in proportion progressing
  • The few earlier events will be either
    • deaths in absence of progression
    • unscheduled scans, probably prompted by deterioration in symptoms
  • In this case, effect on PFS was immediate
Emerging OS data generally support delayed effect

Nivo - melanoma  Nivo – NSCLC Sq  Nivo – NSCLC NSq  Nivo – renal

Pembro – Melanoma  Pembro – NSCLC  Atezol – NSCLC
Non-Proportional Hazards – So What?

• The hazard ratio remains a suitable, primary measure of average effect

However
• The clinical data support presence of delayed effect
• Need supplemental (not replacement) measures of average absolute benefit
  Ones that include all the data recorded from patients
• Important implications for design
• Demonstration of cure would be a game-changer
  Associated statistical challenges

• The excitement about the class is justified
Confidentiality Notice

This file is private and may contain confidential and proprietary information. If you have received this file in error, please notify us and remove it from your system and note that you must not copy, distribute or take any action in reliance on it. Any unauthorized use or disclosure of the contents of this file is not permitted and may be unlawful. AstraZeneca PLC, 2 Kingdom Street, London, W2 6BD, UK, T: +44(0)20 7604 8000, F: +44 (0)20 7604 8151, www.astrazeneca.com