Survival analysis
Coping with non-proportional hazards in randomized trials

Patrick Royston & Mahesh Parmar
Trials Methodology Conference, Bristol, 5 October 2011
Overview

- Basic framework for designing and analysing an RCT with a time-to-event outcome
- Enter non-proportional hazards (non-PH)
- Detecting non-PH, and its possible effects
- Interpreting a hazard ratio in the presence of non-PH
- A possible way forward
- Comments and conclusions
Basic trial requirements

- Every study report needs a summary measure
- Mean, median, odds ratio, hazard ratio, ...
- ... and its variance, leading to a test
- Report measure with 95% CI and P-value
Time to event outcomes

- Censoring presents challenging problems for the analysis of clinical trials
- If it weren’t for censoring, we would almost certainly always work in the *time domain*
  - time to event as a continuous outcome
  - summarize using mean, SD, etc.
  - visualize using scatter plots, etc.
  - analyse using linear regression
- Instead, we have to work in the *probability domain ...* with major implications
Time domain vs. probability domain
(RE01 trial in advanced kidney cancer)

Time domain plot
(ignores censoring!)

Probability domain plot
(allows for censoring)
Working in the probability domain

• For ages, we have used the hazard ratio (HR) as the measure and the logrank statistic as the test
• For the HR to make sense, we have to impose the proportional hazards (PH) assumption
• PH assumption says
  \[ h_1(t) = \text{HR} \times h_0(t) \]
• The survival curves are then related by
  \[ S_1(t) = S_0(t)^{\text{HR}} \]
• This relationship is hard to visualize (i.e. understand)!
Simulated trial to illustrate PH (1)

Weibull distribution in each arm, HR = 0.7
(poor survival)
Simulated trial to illustrate PH (2)

Weibull distribution in each arm, HR = 0.7
(good survival)
So things are fine then ...

- We have a **measure**: the HR
  - we can easily get a CI
- We have a **test**: the logrank (or equivalent Cox) test
- We can live with the **cost** of this framework: the PH assumption
  - It probably isn’t critical anyway – our trials seem to show approximate proportionality
- So everything is fine then ... ?
... or is it? Enter the IPASS trial ...

Crossing curves!

<table>
<thead>
<tr>
<th></th>
<th>gefitinib</th>
<th>gefitinib</th>
<th>gefitinib</th>
<th>gefitinib</th>
<th>gefitinib</th>
<th>gefitinib</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. at Risk</td>
<td>609</td>
<td>363</td>
<td>212</td>
<td>76</td>
<td>24</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>608</td>
<td>412</td>
<td>118</td>
<td>22</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma

Tony S. Mok, M.D., Yi-Long Wu, M.D., F.A.C.S., Sumitra Thongprasert, M.D., Chih-Hsin Yang, M.D., Ph.D., Da-Tong Chu, M.D., Nagahiro Saijo, M.D., Ph.D., Patrapim Sunpaweravong, M.D., Baohui Han, M.D., Benjamín Margono, M.D., Ph.D., F.C.C.P., Yukito Ichinose, M.D., Yutaka Nishiwaki, M.D., Ph.D., Yuichiro Ohe, M.D., Ph.D., Jin-Ji Yang, M.D., Busyamas Chewaskulyong, M.D., Haiyi Jiang, M.D., Emma L. Duffield, M.Sc., Claire L. Watkins, M.Sc., Alison A. Armour, F.R.C.R., and Masahiro Fukuoka, M.D., Ph.D.
Extreme non-PH shock!

- Paper stimulated letters to the journal
- Complaining about inappropriate use of HR and logrank test
- Raises hitherto neglected questions:
  - Meaning of HR when survival curves cross?
  - How *should* we test the null hypothesis and summarize difference between curves?
- In general, *how do we deal with non-proportional hazards?*
Detecting non-PH

- **Graphical**
  - log(-log S) plot against time
  - scatterplot smooth of Schoenfeld residuals for treatment effect against time
  - time-dependent HR in periods (Cox model)
- **Analytical**
  - Grambsch-Therneau (1994) test of trend in Schoenfeld residuals
  - Test time-dependent treatment effect in extended Cox model
A simulated example based on IPASS

Kaplan-Meier survival estimates

trt = CP  trt = Gefinitib
Investigating non-PH

**log(-log S) plot**

- Log(-log survival prob) vs. Log months since randomization
- Smoothed Schoenfeld residuals

**Cox model in 4 periods**

- HR vs. Months since randomization

**Cox model in 25 periods**

- HR vs. Months since randomization
True HR is strongly time-dependent
How can we interpret the HR under non-PH?

- HR depends on length of follow-up time
- In the simulated IPASS example, the HR ranges between 0.2 and 35
- The overall HR is 0.73 (0.65, 0.82) – similar to the result in the published trial
- What does this overall HR mean?
- Some people interpret it as the average HR
- They have proposed methods to estimate the average HR in this situation
  - but are they helpful?
Schemper (2009), citing Therneau & Grambsch (1994), says that “the Cox [model] hazard ratio can be interpreted as an average [HR] over the observed death times”.

Schemper notes that here “all times are taken as equally important, regardless of the different numbers of individuals at risk.”

Schemper goes on to consider several weighted estimates of the average HR.

To us, none of them is intuitively appealing.
Average HR (2)

Schemper (2009): three definitions of a weighted average HR

\[
sAHR = \int \frac{h_1(t)}{h_0(t)} w(t)f(t)dt
\]

\[
gAHR = \exp \left[ \int \log \left( \frac{h_1(t)}{h_0(t)} \right) w(t)f(t)dt \right]
\]

\[
AHR = \frac{\int (h_1(t)/(h_0(t) + h_1(t))) w(t)f(t)dt}{\int (h_0(t)/(h_0(t) + h_1(t))) w(t)f(t)dt}
\]
Average HR (3)

• We think the average HR is not uniquely defined and not readily interpretable
• We therefore don’t recommend it
A possible way forward
Restricted mean survival time (RMST)

- **Motivation:**
  - We choose to work in the time domain
  - We want the *mean* survival time
  - But because of censoring, we do not observe the entire survival distribution
- Select a time point, \( t^* \), up to which we wish to compute the *restricted mean survival time*
- Formally, for a random time-to-event variable \( T \), we estimate
  \[
  \mu(t^*) = E(\min(T, t^*)) = \int_0^{t^*} S(t) dt
  \]
Interpretation of RMST

• RMST = area under the survival curve up to $t^*$
• Can think of it as the ‘$t^*$-year life expectancy’
• A patient might be told that ‘your life expectancy with Z disease on X treatment over the next 18 months is 9 months’
• Or, ‘treatment A increases your life expectancy during the next 18 months by 2 months, compared with treatment B’
Estimation of RMST

The idea of RMST goes back to Irwin (1949)

There are several methods for estimating it:

- Integrate Kaplan-Meier survival curves
- Jackknife (Andersen et al 2004)
- Flexible parametric models (Royston & Parmar 2002)
- Others ...
- See Royston & Parmar (2011) for details
Choice of $t^*$

- Should be pre-specified
- Should be chosen to cover the follow-up period of clinical interest
- $t^*$ is an important aspect of a pre-specified statistical analysis plan
- Usually close to the last observed death time
A recent MRC CTU example: ICON7

- RMST difference at t* = 19 mths is 1.5 (0.2-2.9) mths
- HR = 0.81 (0.70-0.94)
A suggested analysis strategy
(Royston & Parmar 2011, sec. 4)

- Apply logrank test as usual
- Test the treatment effect for non-PH
- If PH, primary summary is the HR and CI
- If non-PH, primary summary is the difference in RMST at $t^*$ and CI
- Even when PH looks OK, RMST is still a useful measure that can be presented along with other measures and analyses
Further work is in progress on RMST

- When PH is true, power of RMST difference seems to be comparable to logrank test
- When non-PH is present, power comparison is more complicated
- If expecting non-PH, may design a trial with RMST difference as the target measure
Conclusions and comments

- It’s poor practice to calculate and report a single hazard ratio when non-proportional hazards is clearly present.
- RMST is an attractive alternative measure, and is easy to compute.
- RMST can be calculated for any model for the data – it’s as general as a Kaplan-Meier curve.
- For now, we should still continue to design trials with a single hazard ratio in mind:
  - Logrank test retains reasonable power even when moderate non-PH is present.
Some references


Thank you.