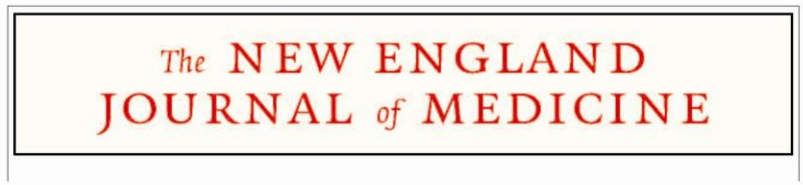
Analysis and design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated

Patrick Royston & Mahesh Parmar
MRC CTU

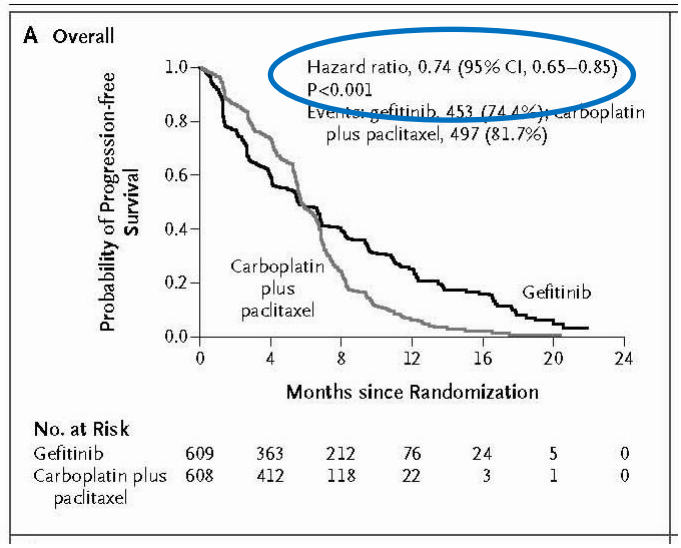
NEJM 2009;361



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., Benjamin 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.

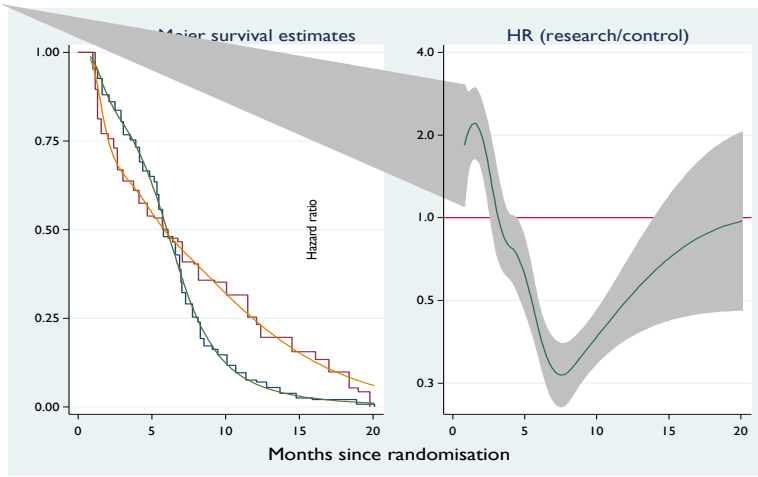
… non-ignorable non-PH does happen!



Overview

- Recent trials have exhibited noticeable **non-PH**
  - E.g. IPASS, ICON6, ICON7 …
- Means that the treatment effect **depends on time**
  - Important for interpretation, analysis and design
- Logrank/Cox test may be **severely underpowered**
- We badly need a **more robust** test
- Key idea: restricted mean survival time (**RMST**)
- Develop **RMST-based tests** of the treatment effect
- Combine with logrank/Cox to get best of both tests
- Investigate power of tests
- An approach to **robust trial design**
- Conclusions

IPASS (recon): PFS KM and HR vs. time



Survival curves estimated using a flexible parametric model, PH(5,5)

What does the overall hazard ratio mean?

- In the reconstructed IPASS example, the HR ranges between 0.27 and 2.2 over time
- The overall HR at the time of this analysis is 0.73 (95% CI 0.64, 0.83)
- What does this mean?
- Some people (e.g. Schemper 2009) have interpreted the overall HR as a type of a weighted average HR over the event times
- But we think a single HR when there is non-PH is not interpretable
- Instead we work with RMST

Restricted mean survival time (RMST)

- Suppose we have a set of observed and censored time-to-event data
- Motivation: it's natural to summarize through the mean, but we can't because we haven't observed the entire survival distribution
- Select a time point, t^*, up to which we wish to compute the *restricted* mean survival time

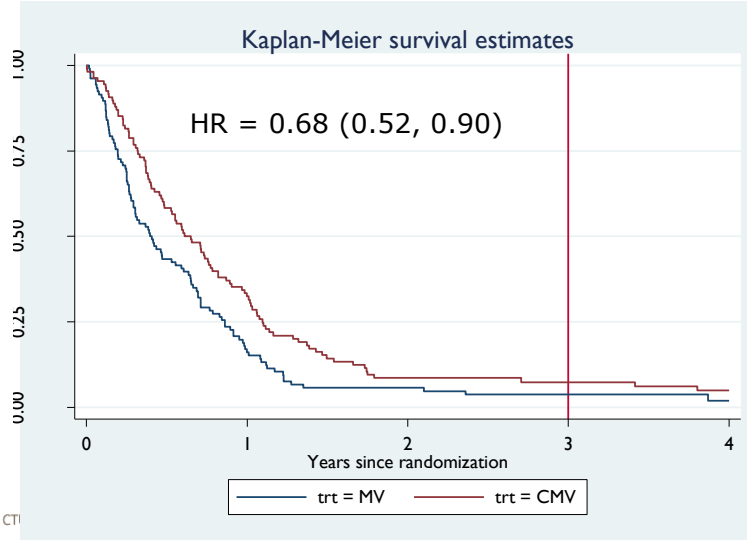
Interpretation of RMST

- 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*'

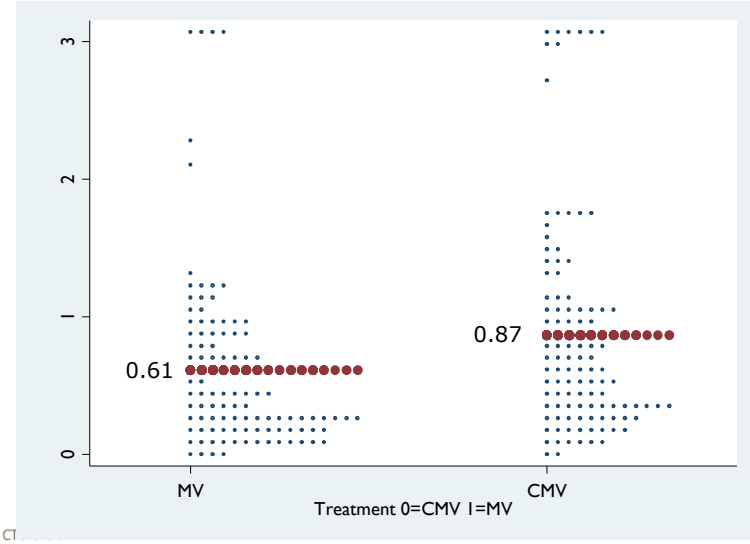
Estimation of RMST

- The idea of RMST goes back to Irwin (1949) and Kaplan & Meier (1958)
- There are several methods for estimating it:
  - Non-parametric (Kaplan-Meier survival curve)
  - Jackknife (Andersen et al 2004)
  - From flexible parametric models (Royston & Parmar 2002, Royston & Lambert 2011)
- In Stata there is `predict` after `stpm2`, and `strmst` specifically for trials data; also `stpmean` for jackknife estimation
  - All user-written

Example: BA07 in advanced bladder cancer



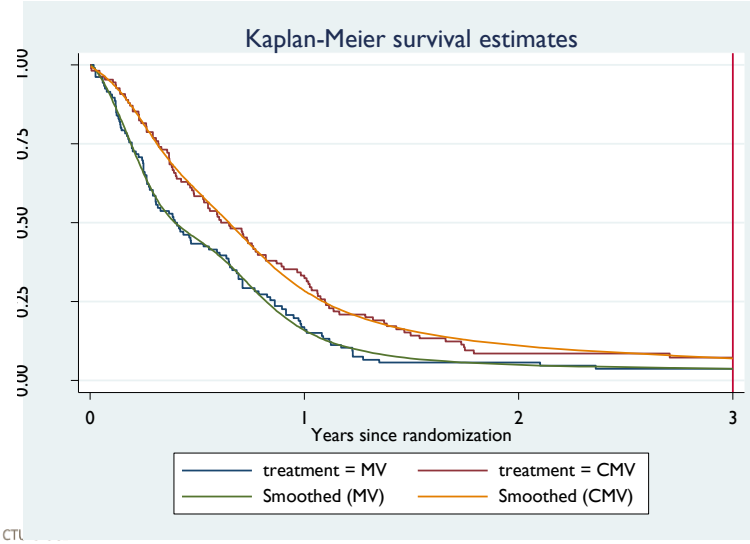
Jackknife RMST estimates for individuals (t* = 3 years) using `stpmean`



Flexible parametric models

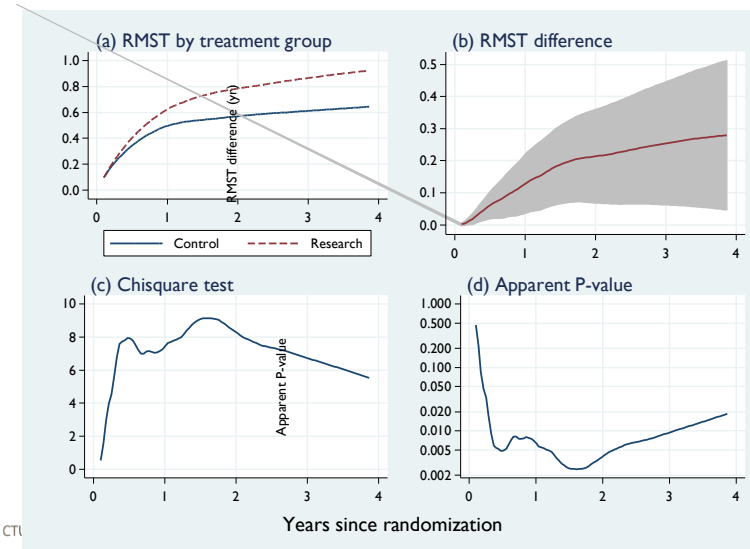
- Royston & Parmar (2002), Royston & Lambert (2011), Lambert & Royston (2009)
- Estimates baseline cumulative hazard function as a smooth (spline) function of time *t*
- Gives (smooth) estimates of *S*(*t*) etc etc
  - Less noisy than Kaplan-Meier estimates
- Can include time-dependent treatment effects
- Can include covariate effects
- Stata program `stpm2`

BA07 again: smoothing by FPM



MRC CTU

Time-dependent RMST treatment plots for BA07



MRC CTU

Test for treatment effect based on RMST?

- Treatment effect: ΔRMST = difference in RMST between two trial arms
  - Research minus control
- ΔRMST and its P-value depend on t*
- Choosing a fixed t* is a poor strategy
- A better approach: find the smallest P-value for ΔRMST over a sensible range of t* values
- In BA07, P = 0.0025 at t* = 1.47 years
- But this P-value is obviously "too small"
  - Multiple testing
- What can we do about this?

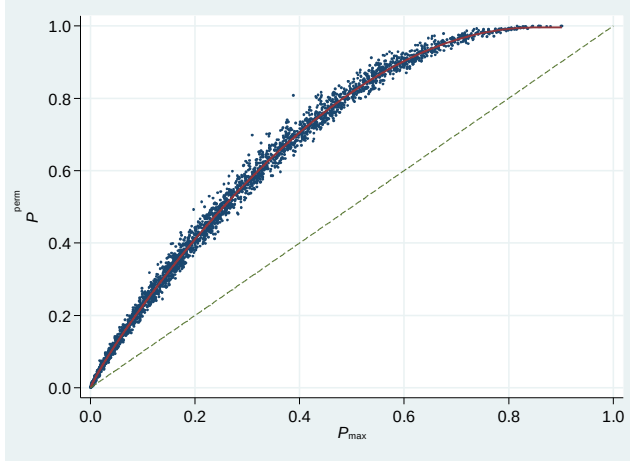
A permutation test for RMST difference

- Aim is to correct the minimal P-value from ΔRMST for multiple testing
- Randomly permute the treatment label many times
- In each permuted sample, compute the RMST-based minimal P-value
  - This estimates the "null distribution" of the P-value
- Determine relative rank position of original P-value in the "null distribution" of the P-value
- This gives the permutation test P-value

Approximating the permutation test

- Permutation test is a reasonable approach
- But it has drawbacks
  - Random element means the test is not exactly reproducible from run to run
  - Test is slow to compute
- An approximation to the test is helpful
  - Derived from the uncorrected P-value
  - Simulation in null case (no treatment effect)

Permutation P-value versus original P-value



- Data simulated from 3 trials x 1000 replicates
- Red line shows a suitable regression model fit

Example: Bladder (BA07) trial

- Cox test: P = 0.0069
- RMST original P-value: P = 0.0025
- Permutation test (9999): P = 0.0089 (0.0073, 0.0110)
- Approximate perm. test: P = 0.0087

The Royston-Parmar (RP) test (1)

- Under some non-PH scenarios the permutation test has higher power than the Cox/logrank test
- Under PH and (some) non-PH scenarios, the permutation test has lower power than the logrank/Cox test
- Can we get the **best** out of Cox and permutation tests?
- Aim: get good power for PH and non-PH scenarios
  - i.e. create a more robust test
- Approach: Royston-Parmar (RP) test, aka combined test

- Key idea: take the **smaller** of the P-values from the Cox and approximate permutation tests
- Compute $P_{min} = \min(P_{Cox}, P_{perm})$

The Royston-Parmar (RP) test (2)

- Need to adjust P_{min} because it is the smaller of two P-values
- Call the adjusted P_{min} P_{RP}
- Estimate P_{RP} using simulation based on several trial datasets
- Derive an empirical approximation using a beta distribution:
  - P_{RP} = ibeta(P_{min} ; 1, 1.5)
  - For small P_{min}, $P_{RP} \approx P_{min} \times 1.5$
- Have gained something compared with Bonferroni correction, since Bonferroni would give $P_{Bon} = P_{min} \times 2$

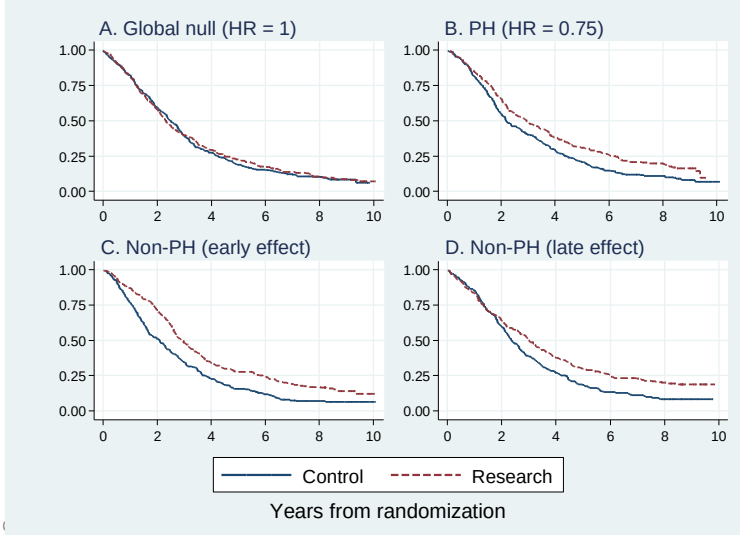
Examples

- $P_{min} = 0.05$, $P_{RP} = 0.074$
- $P_{min} = 0.0336$, $P_{RP} = 0.05$

The Royston-Parmar (RP) test (3): scenarios used in simulations of power

- A. PH (HR = 1): null case
- B. PH (HR = 0.75)
  - Quite common
  - Often reasonable in trials with short follow-up time
- C. Non-PH (early effect)
  - HR starts <1 and approaches or exceeds 1 over time
  - Reasonably common
  - E.g. in trials with differently acting treatments
- D. Non-PH (late effect)
  - HR starts ~1 for a period then reduces over time
  - Less common but not rare
  - May occur in screening or prevention trials

Examples of scenarios (simulated data)



MRC C

Simulation results: type 1 error and power

Scenario	Dataset	*n*	Test		
			Cox	Perm.	RP
A (null)	GOG111	1000	5.2	4.9	5.3
	PATCH1	1000	5.0	5.2	4.8
	ICON7	1000	4.8	5.2	4.9
B (PH)	GOG111	652	**92.9**	86.7	91.0
	PATCH1	1280	**92.6**	87.7	90.2
	ICON7	1240	**91.9**	88.3	89.8
C (early)	GOG111	310	72.5	**92.1**	90.0
	PATCH1	450	74.4	**91.9**	89.2
	ICON7	522	36.9	92.4	89.5
D (late)	GOG111	560	92.7	80.5	90.3

Benchmark: ~90% power for RP test

Example: Robust design based on RP test

- Based on advanced bladder cancer (BA06)
  - Estimate or assume control arm survival function
  - Recruit 3 years, follow up 3 years
- Logrank/Cox test under HR = 0.75 for power 90% at significance level 5% requires 796 patients (509 events)
- RP test under HR = 0.75 for power 90% at significance level 5% requires 851 patients (544 events)
- The "insurance premium" needed for the RP test is about 100 x (851-796)/796 = 7% in this example
- Aims to protect power under many non-PH scenarios
  - Particularly, treatments with "early effect"

Software

- Stata
- `strmst` performs the RP test: submitted to *Stata J*
- `stpower_rp` performs power and sample size calculation for the RP test: under development, version 1.0 done
- `stpower_rp` also plots population survival curves based on your specified control survival and HR functions
- Ask PR if you would like to try out these packages (and give comments, if possible)

Conclusions

- Difficult to predict whether non-PH will be present or not
- Even if you suspect it will, what shape will the HR function take over time? Unclear.
- Use of restricted mean survival time facilitates testing and displaying a generalized treatment effect
- The RP test increases trial power under an early treatment effect and protects power under several other scenarios
- RP test requires an "insurance premium" of <10% increase in sample size

Some references

- Andersen PK, Hansen MG & Klein JP. Regression analysis of restricted mean survival time based on pseudo-observations. *Lifetime Data Analysis 2004*; 10: 335-350.
- Irwin JO. The standard error of an estimate of expectation of life, with special reference to expectation of tumourless life in experiments with mice. *Journal of Hygiene* 1949; 47: 188–189.
- Overgaard M, Andersen PK & Parner ET. Regression analysis of censored data using pseudo-observations: an update. *Stata Journal* 2015; 15: 809-821.
- Royston P & Lambert PC. *Flexible parametric survival analysis using Stata: Beyond the Cox model*. StataPress, TX, 2011.
- Royston P & Parmar MKB. 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* 2011; 30:2409-2421.
- Royston P & Parmar MKB. Augmenting the logrank test in the design of clinical trials in which non-proportional hazards of the treatment effect may be anticipated. *BMC Medical Research Methodology* 2016; 16:16