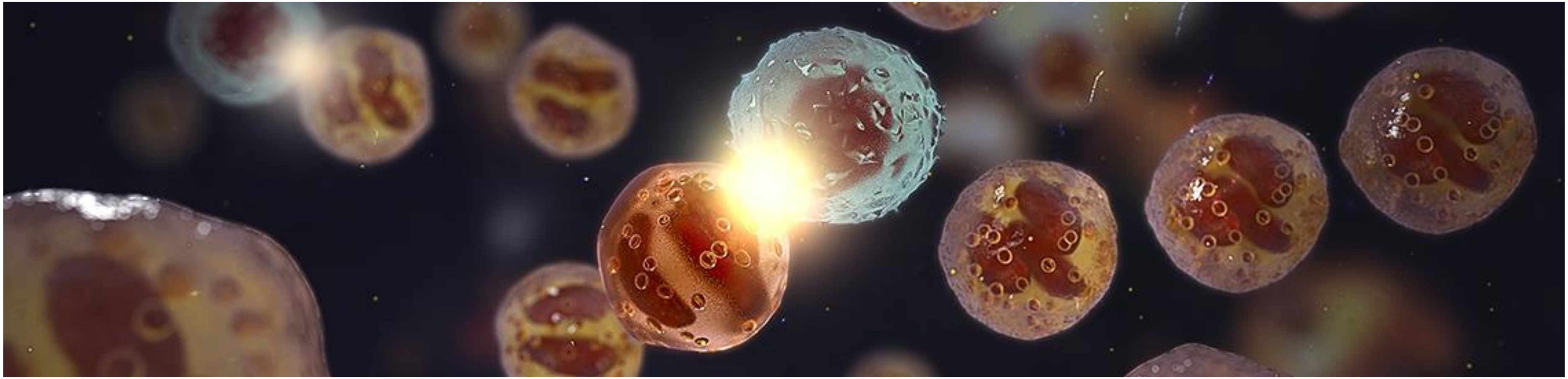
Non-Proportional Hazards – So What?

Andrew Stone

BBS spring seminar April 28th 2016



Disclaimer

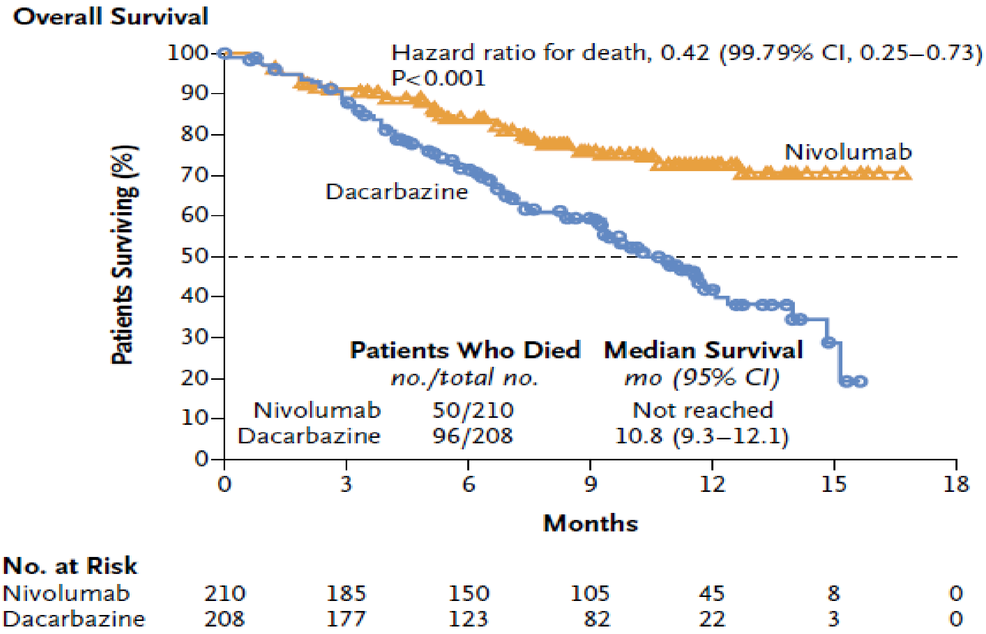
Andrew Stone is an employee of AstraZeneca LP. The views and opinions expressed herein are my own and cannot and should not necessarily be construed to represent those of AstraZeneca or its affiliates.

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



	Patients Who Died *no./total no.*	Median Survival *mo (95% CI)*
Nivolumab	50/210	Not reached
Dacarbazine	96/208	10.8 (9.3–12.1)

No. at Risk							
Nivolumab	210	185	150	105	45	8	0
Dacarbazine	208	177	123	82	22	3	0

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Nivolumab in Previously Untreated Melanoma without *BRAF* Mutation

C Robert et al. N Engl J Med 2015;372:320-330.

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 (ie, 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.’*[6]

Really?? Let’s examine this assertion.

* Uno H, J Clin Oncology 2014 2380-5

Is this any different to multicentre clinical trials?

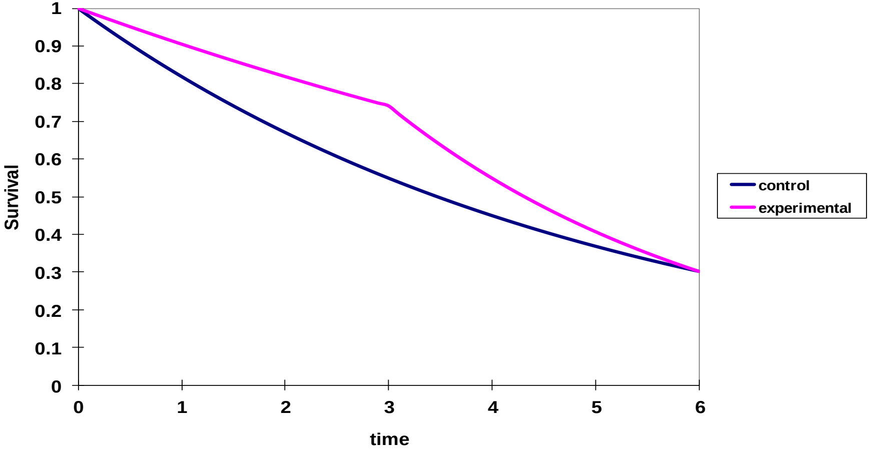
- Quantitative treatment-by-centre interaction
  - We're quite happy to describe the treatment benefit as the average over centres even if there is statistical evidence that the benefit differs across centres

- 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

Time Period	Hazard rate		HR
	New	Control	
0-3	0.1	0.2	0.5
3-6	0.3	0.2	1.5
HR estimated by Cox			0.86
Weighted average of piecewise HR			0.86



- 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

For future reference: why.

ln(HR) ~ U/V*,

- where U = $\sum_{i}(d_{1j} - {}^{n_{1j}d_j}\!/_{n_j})$ the usual log-rank denominator
- and V = $\sum_{i}\frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)}$ ~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

$$=(U_1+U_2)/(V_1+V_2)$$
$$= (e_1/4 . \ln(HR_1) + e_2/4.\ln(HR_2)) / (e_1/4 + e_2/4)$$
$$= p_1 \ln(HR_1) + p2 \ln(HR_2)$$

** Berry G, Kitchin RM, Mock PA. A comparison of two simple hazard ratio estimators based on the logrank test. Statistics in Medicine 1991; 10:749-755*

Sellke, T. and Siegmund, D. Sequential analysis of the proportional hazards model. Biometrika 70: 315-326, 1983

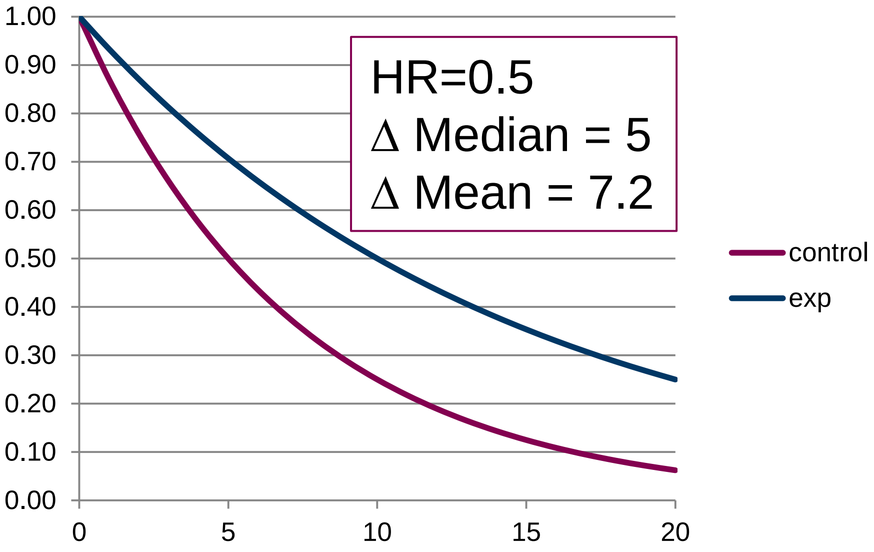
Ah but...

- 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

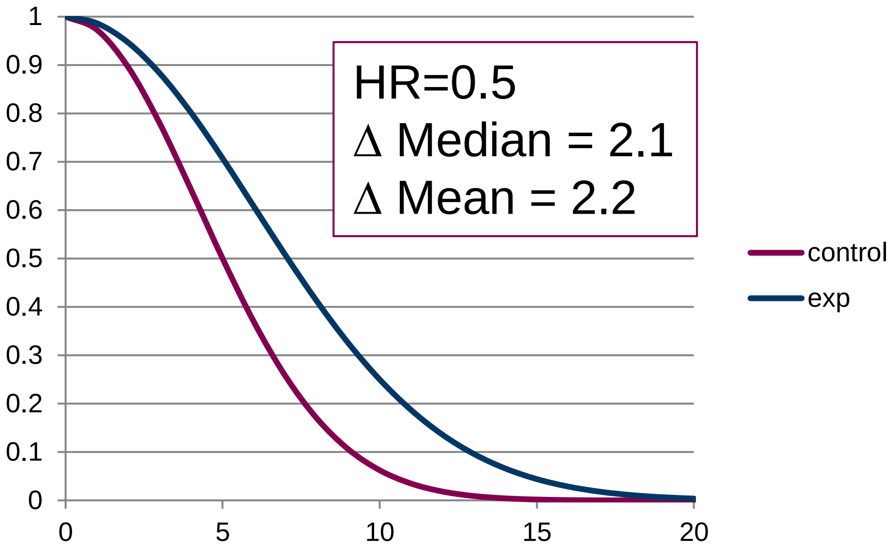
- But should acknowledge in interpretation that with further follow-up that treatment effect will change
  - Importance of follow-up
  - This applies equally to the alternatives proposed

However, regardless of proportionality ....



HR=0.5
Δ Median = 5
Δ Mean = 7.2

Exponential distn
Control median = 5



HR=0.5
Δ Median = 2.1
Δ Mean = 2.2

Weibull distn $S(t) = \exp(-0.028*t^2)$
Control median = 5

Same meaning clinically?

We need to supplement with absolute benefit

NOTE: A kaplan-meier of the ranks looks identical for both distibutions, as HR based on relative ranking not actual times

HR remains meaningful and the primary measure of effect

But supplemental measures needed

But what?

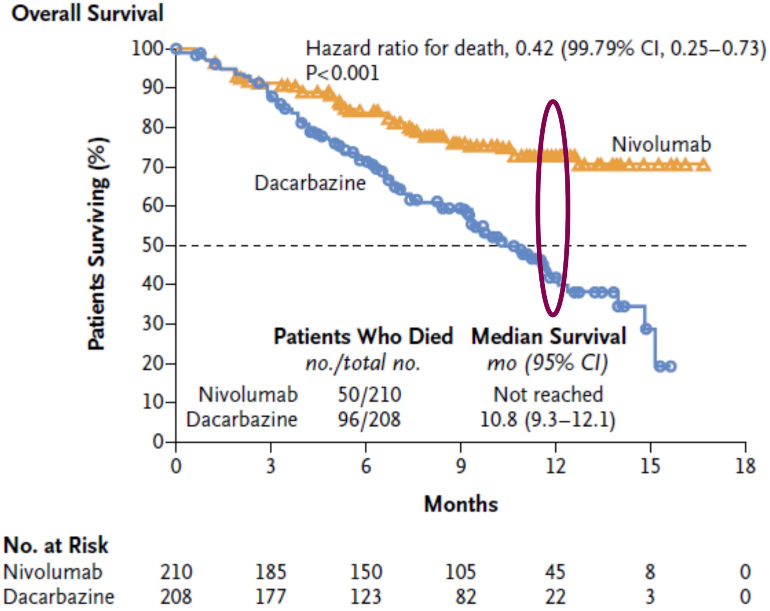
Medians normally used for absolute benefit, yet we know they're a lousy measure

- 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]



	Patients Who Died *no./total no.*	Median Survival *mo (95% CI)*
Nivolumab	50/210	Not reached
Dacarbazine	96/208	10.8 (9.3–12.1)

No. at Risk

Nivolumab	210	185	150	105	45	8	0
Dacarbazine	208	177	123	82	22	3	0

[1] *Klein JP, et al. Analyzing survival curves at a fixed point in time. Stat Med 2007;26:4505-4519.*

[2] *Stone A et al. Improving the design of Phase II trials of cytostatic anti-cancer agents. Cont Clin Trials 2007 28: 138-145*

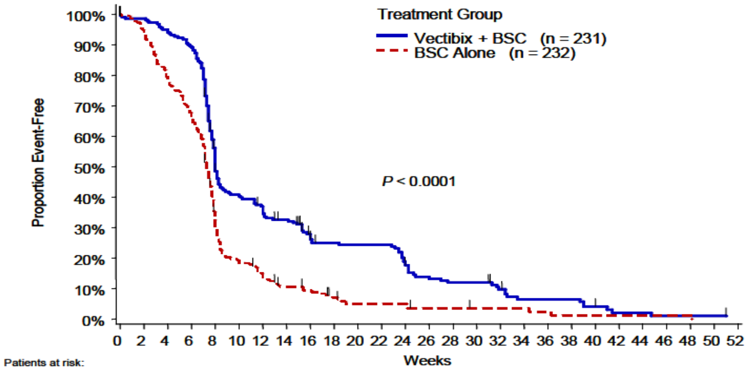
What about mean survival

- 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



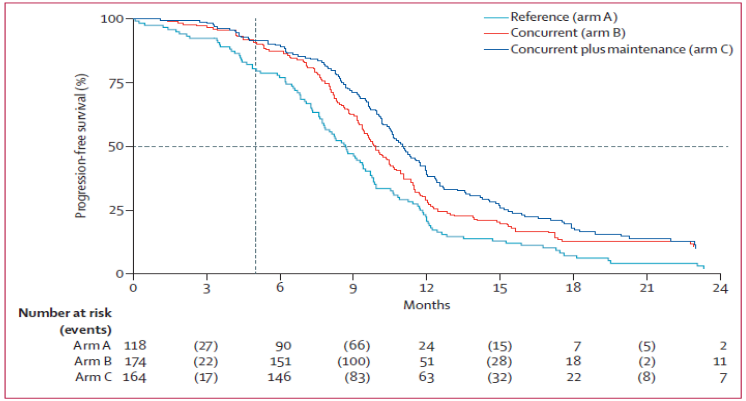
Patients at risk:																										
Vectibix + BSC	231	217	209	197	118	85	76	65	49	41	40	40	31	22	19	19	13	8	8	8	5	2	2	1	1	1
BSC Alone	232	209	175	149	75	41	31	20	17	11	7	7	7	4	4	3	3	3	2	1	1	1	1	1	1	0

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, UK[1]
- 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



Number at risk (events)									
Arm A	118	(27)	90	(66)	24	(15)	7	(5)	2
Arm B	174	(22)	151	(100)	51	(28)	18	(2)	11
Arm C	164	(17)	146	(83)	63	(32)	22	(8)	7

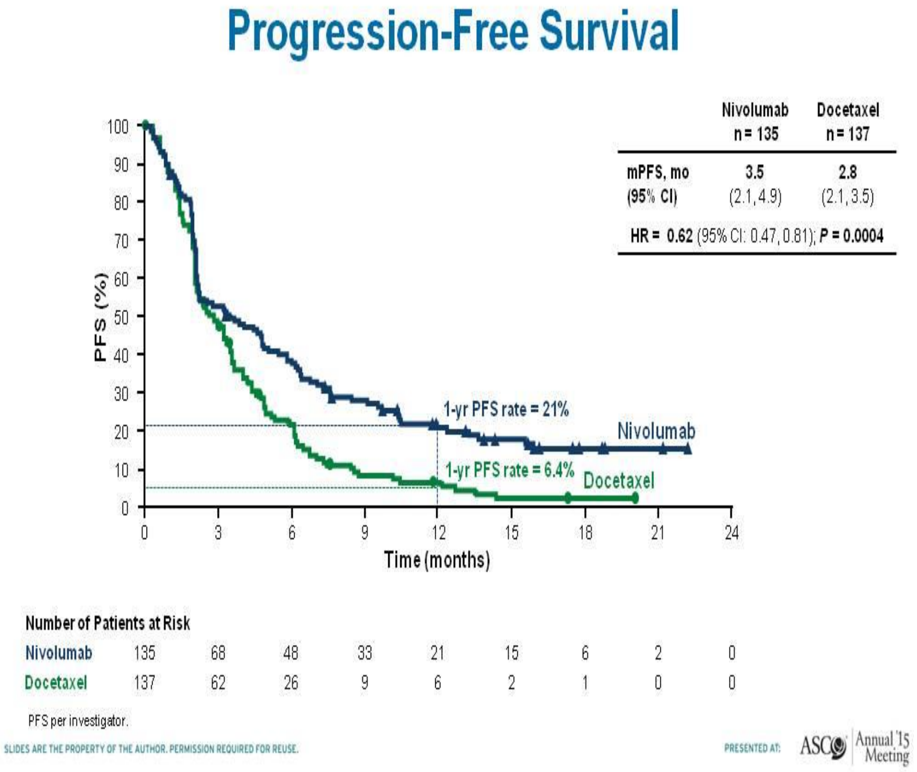
Figure 2**: Kaplan-Meier plot of progression-free survival over 2 years**
Vertical reference line shows the median time to completion of the chemotherapy phase. Number at risk every 6 months shown with the number of failure events in parentheses, after the time in which the number at risk was calculated.

'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]

[1] *Royston and Parmar BMC Medical Research Methodology 2013, 13:152*

[2] *Ledermann et el.. Lancet Vol 387 March 12, 2016*

Mean restricted regardless of period of proportionality



Progression-Free Survival

	Nivolumab n = 135	Docetaxel n = 137
mPFS, mo (95% CI)	3.5 (2.1, 4.9)	2.8 (2.1, 3.5)
HR = 0.62 (95% CI: 0.47, 0.81); *P* = 0.0004		

Number of Patients at Risk

Nivolumab	135	68	48	33	21	15	6	2	0
Docetaxel	137	62	26	9	6	2	1	0	0

PFS per investigator.

SLIDES ARE THE PROPERTY OF THE AUTHOR. PERMISSION REQUIRED FOR REUSE.

PRESENTED AT: ASCO Annual '15 Meeting

Estimated from curves:

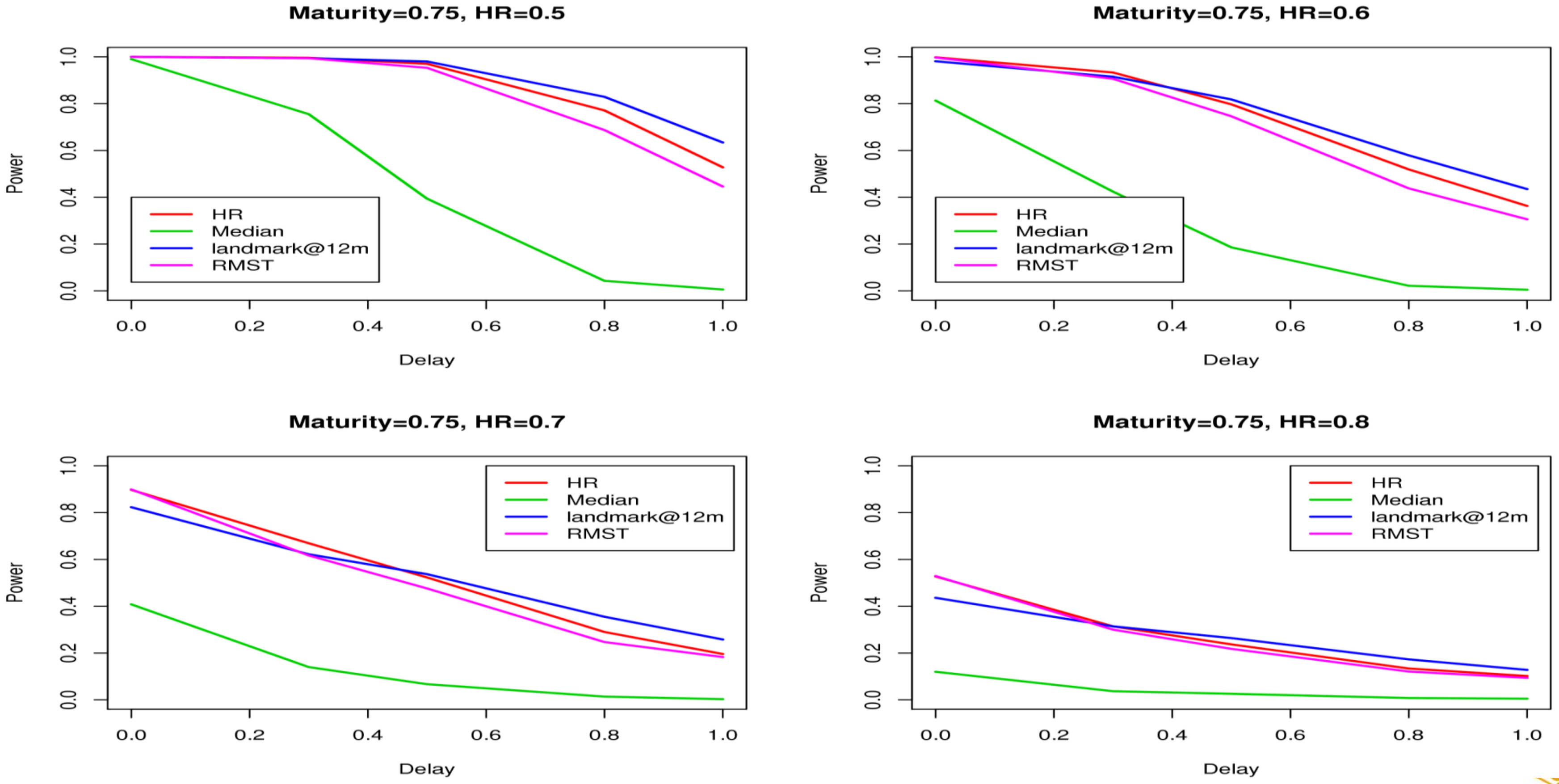
•RMST difference with truncation point at 12 months ~1.5

• 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-axis = T/median where HR=1 < T with T=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

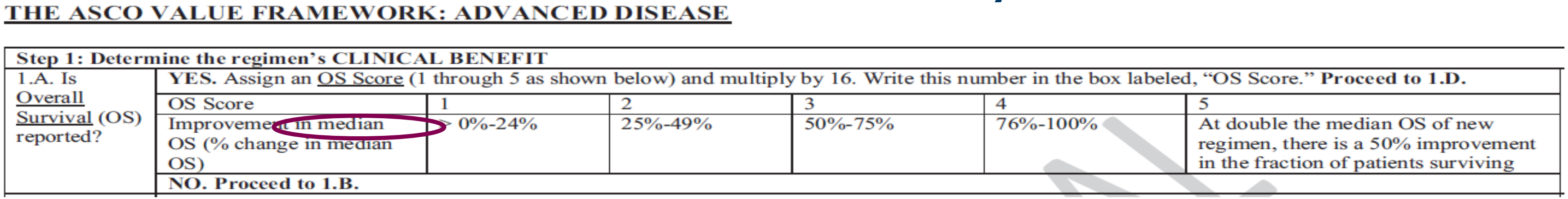
- See Ellis[2] for means and variances
- Advantage: these would represent overall predicted means rather than those restricted to a timeperiod. To be studied further

[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:

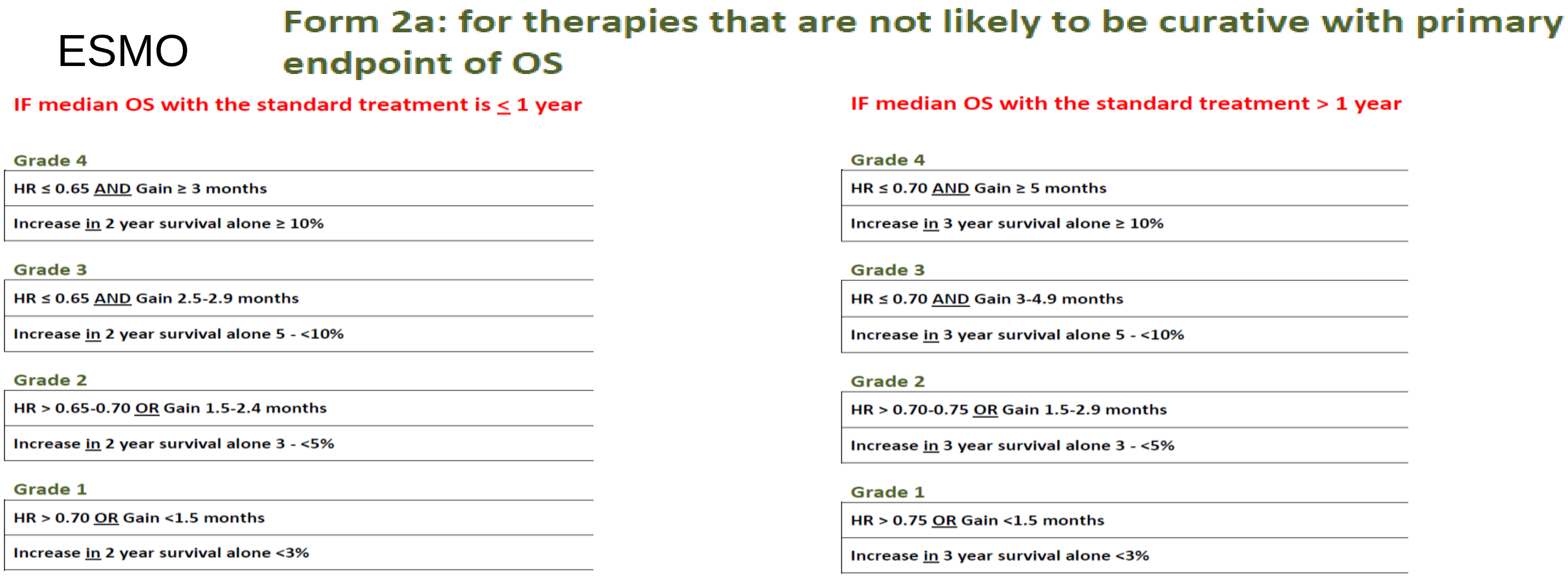
Scenario: Advanced Disease with OS as Primary



THE ASCO VALUE FRAMEWORK: ADVANCED DISEASE

Step 1: Determine the regimen's CLINICAL BENEFIT						
1.A. Is Overall Survival (OS) reported?	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.**					
	OS Score	1	2	3	4	5
	Improvement in median OS (% change in median OS)	> 0%-24%	25%-49%	50%-75%	76%-100%	At double the median OS of new regimen, there is a 50% improvement in the fraction of patients surviving
	NO. Proceed to 1.B.					

ESMO



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
HR_1=1 $t<T$, HR_2 =x (<1) $t \geq T$,
where T denotes the lag-time until there is a benefit of therapy and HR_2 the hazard ratio (experimental : control) before and after the lag respectively.

The overall average HR is given by [1,2]:

$$\overline{HR} = \exp\left(p_2 \ln(HR_2)\right)$$

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:

$$p(event\ by\ time\ t) = \left(\frac{\min(t,B)}{B}\right)^k - \frac{k}{B^k}\int_0^{\min(t,B)} s^{k-1} S(t-s)\, ds$$

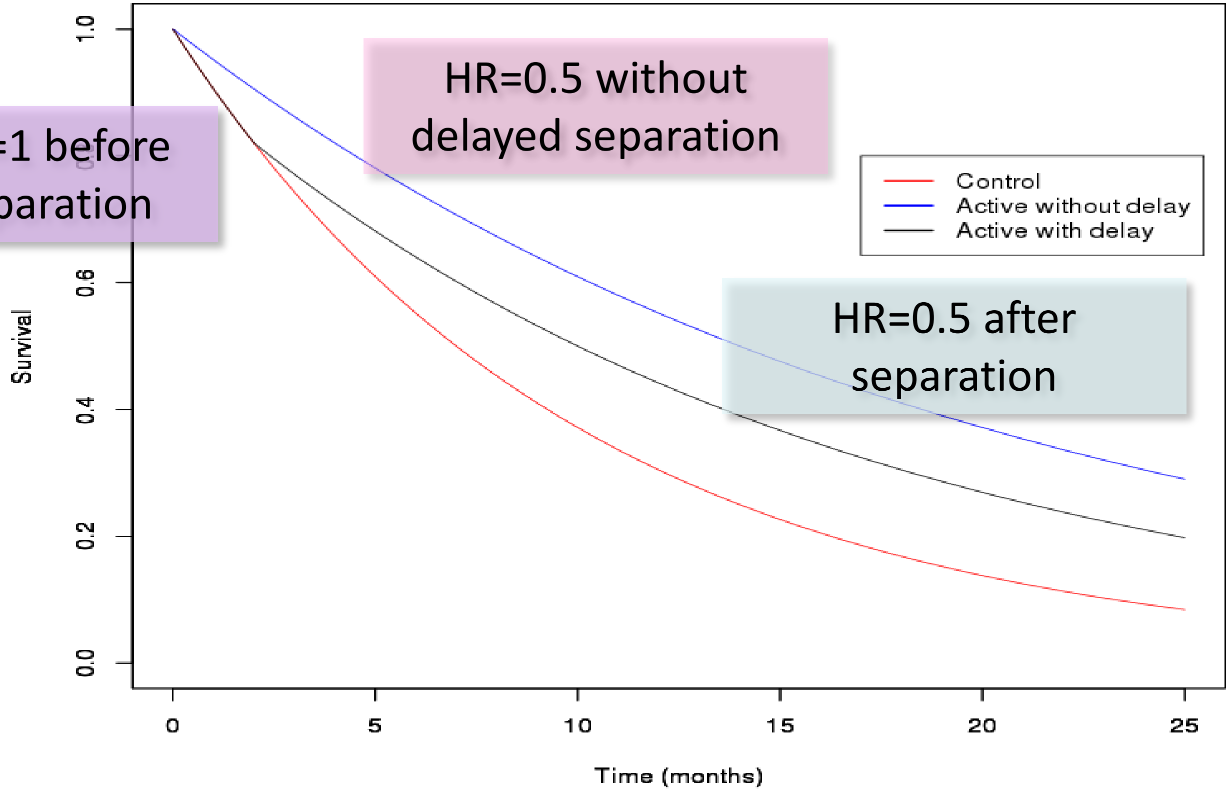
$\overline{HR}$ can then be calculated and together with n, the total no. of events, the following can be re-arranged to estimate power

$$n = \frac{(1+r)^2}{r} * \frac{\left[\Phi^{-1}(1-\alpha/2) + \Phi^{-1}(1-\beta)\right]^2}{ln^2(\overline{HR})}$$

Where r= randomisation ratio (eg r=1 with 1:1 randomisation)

[1] Kalbfleisch, J. D., and Prentice, R. L. (1981), “Estimation of the Average Hazard Ratio”, *Biometrika, 68, 105-112.*
[2] Schemper, M. (1992), “Cox Analysis of Survival Data with Non-Proportional Hazard Functions”, *The Statistician, 41, 455-465.*
[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

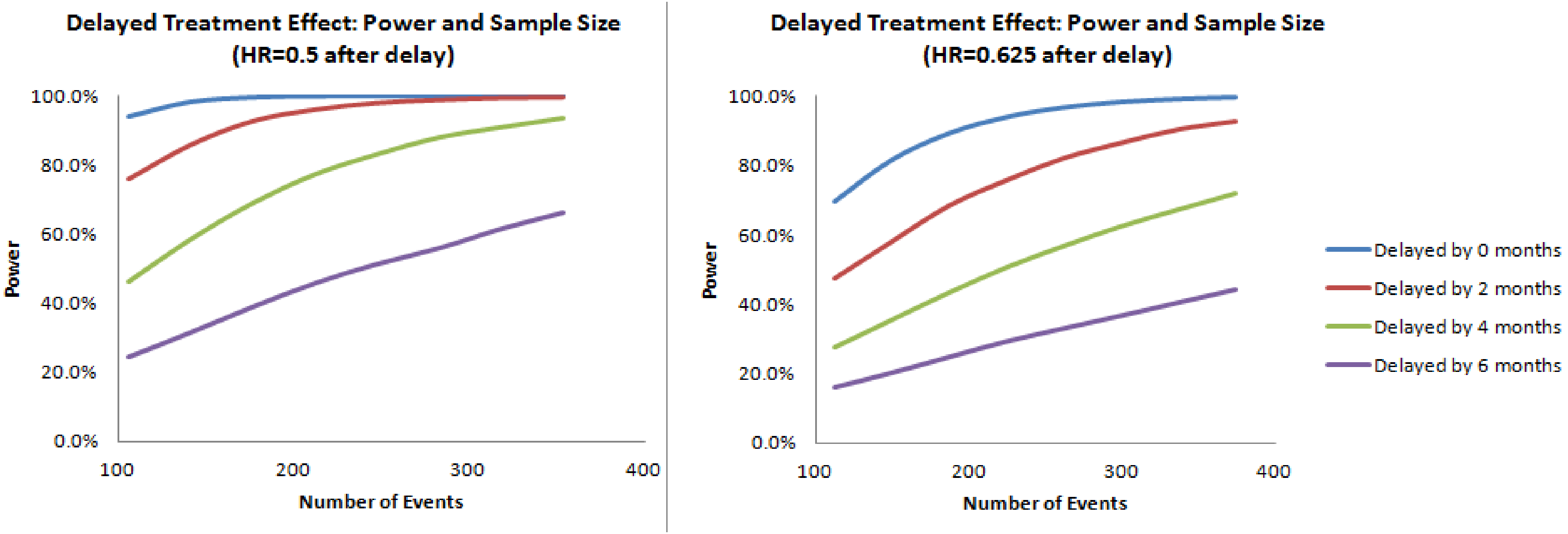


HR=1 before separation

HR=0.5 without delayed separation

HR=0.5 after separation

Adverse impact on power if delay is not accounted for

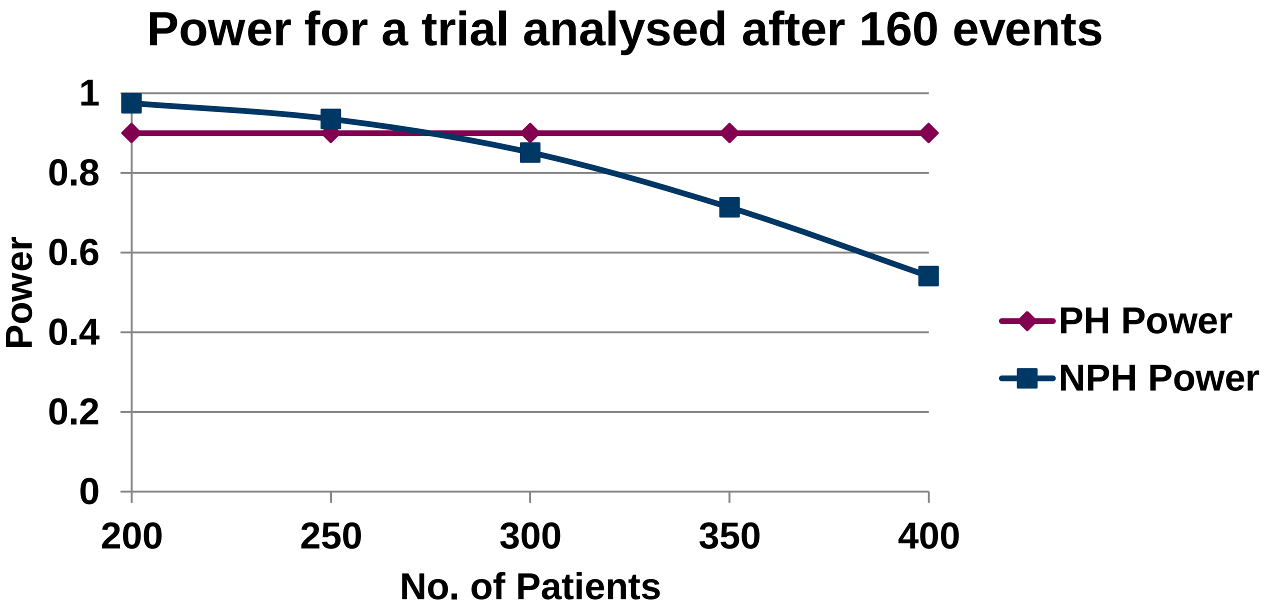


1:1 randomization; assumes e/0.71 patients are recruited where e = no. of events
Fixed 15 months accrual time (uniform);
Median OS (control)=7 months; 2-sided type I error =0.05;

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

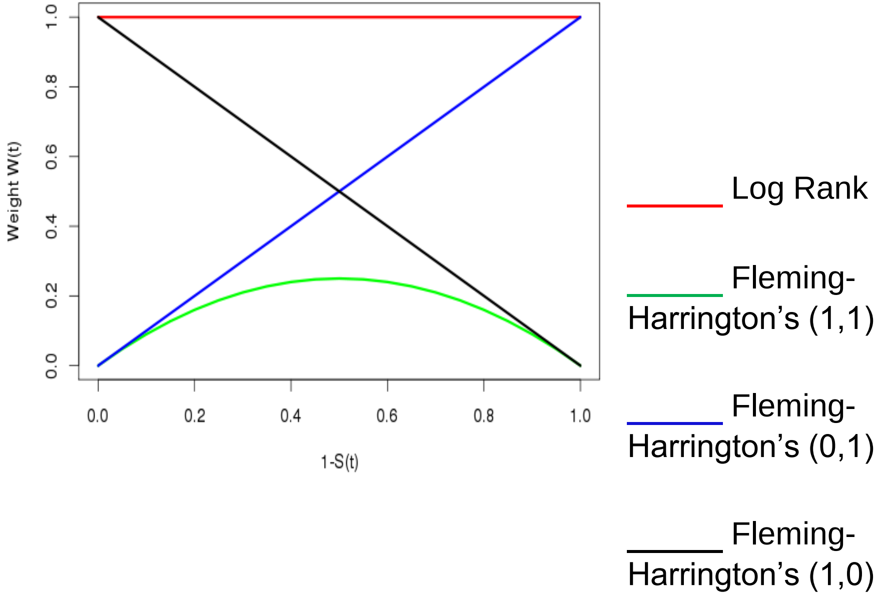
An alternative approach to futility analyses maybe needed

- 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



[1] Fleming, T.R., and Harrington, D.P. (1991), *Counting Processes and Survival Analysis, John Wiley & Sons, New York.*

Yes

	Power				
	Under H0	T=0	T=2	T=4	T=6
Log-rank	**4.8**	**89.9**	**67.5**	**43.3**	**23.5**
$G^{0,1}$	**5.5**	**79.4**	**74.7**	**60.5**	**41.2**
$G^{1,1}$	4.9	85.9	78.1	55.2	29.8
$G^{1,0}$	5.4	85.8	50.9	24.5	12.1

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 α=5%

Simulations by Wenmei Huang

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

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

p =proportion cured and weibull survival in non-cured

- Long term cured model derived based on nivolumab melanoma data

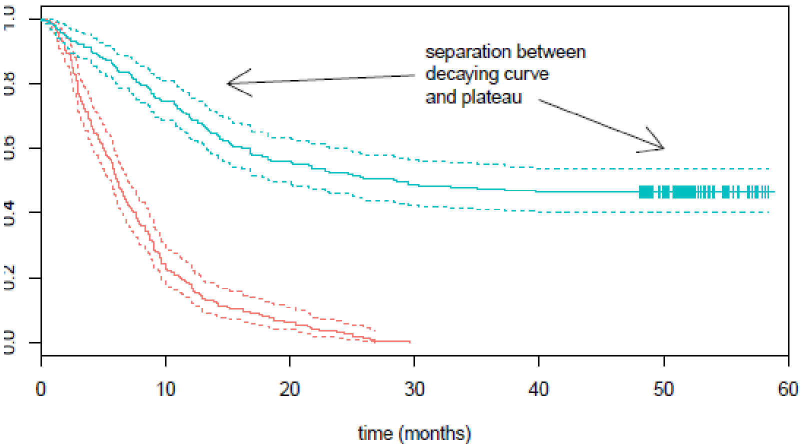
- Two questions investigated

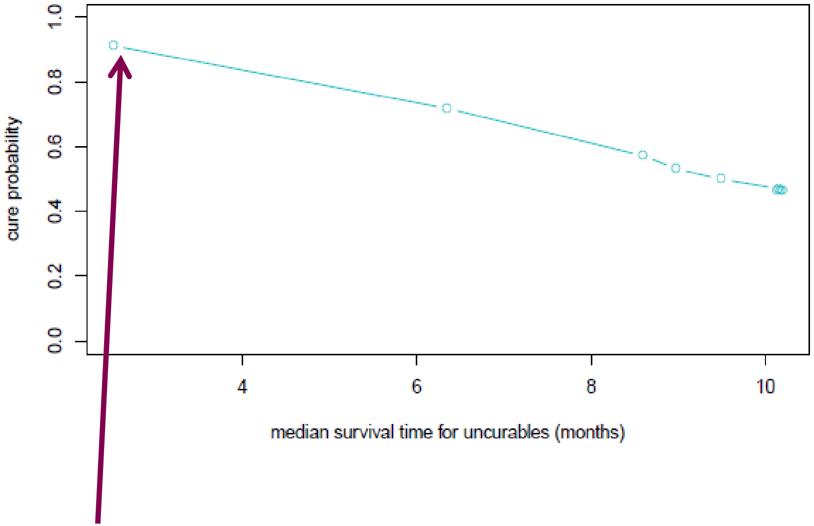
  1) How do you know whether you’ve correctly identified the cure rate?

  2) How much follow-up do you need?

Models will always provide an estimated cure fraction. But not necessarily the correct one!

True curve simulated from



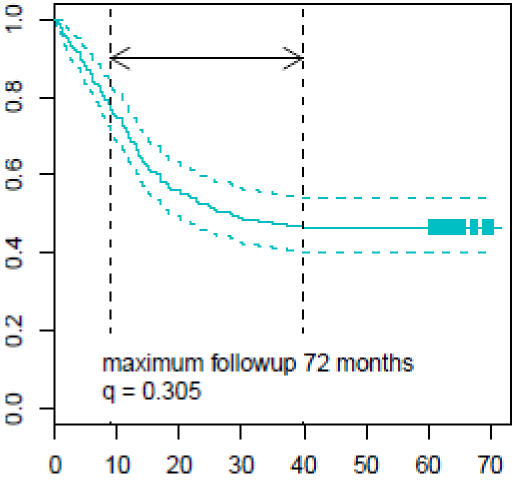


- Cure fraction badly over-estimates truth if analysed too early
  - 95% CI excludes truth (not shown)
- To the extent that lower CI for cure fraction excludes the truth

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*

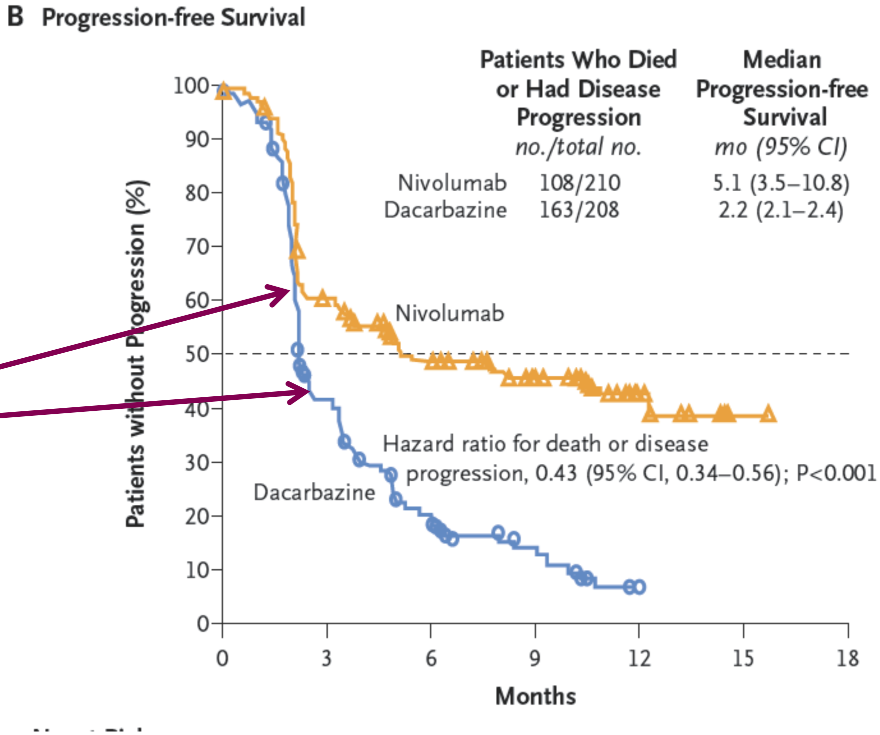
Just finally – what's the emerging picture in terms of presence of delayed effect?

Careful how PFS curves are (mis?) interpreted

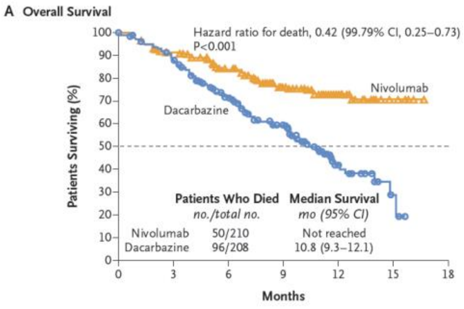
- Care required when examining presence of time-lag with PFS data

- Have seen KM curves misinterpreted a few times

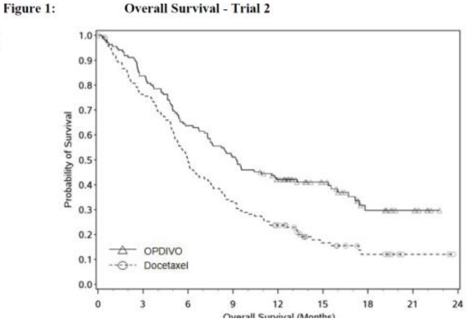
- In the example opposite:
  - At the first scheduled scan, 9 weeks, there is a clear difference in proportion progressing
  - The few earlier events will be either
    - deaths in absence of progression
    - unscheduled scans, probably prompted by deterioration in symptoms
  - **In this case, effect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



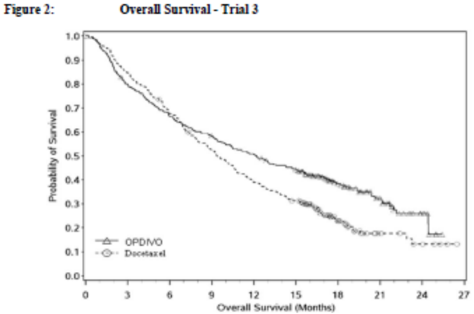
Emerging OS data generally support delayed effect



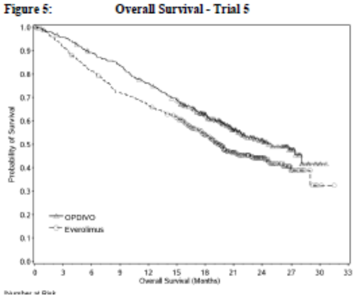
Nivo - melanoma



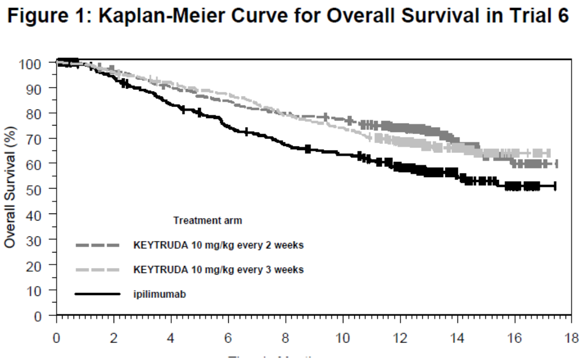
Nivo – NSCLC Sq



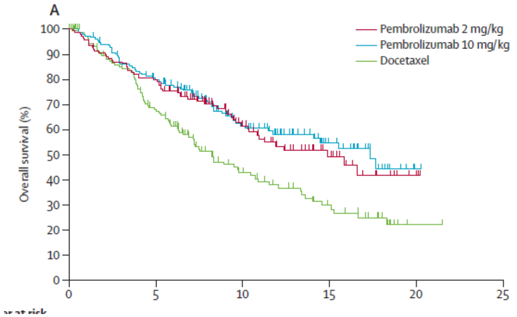
Nivo – NSCLC NSq



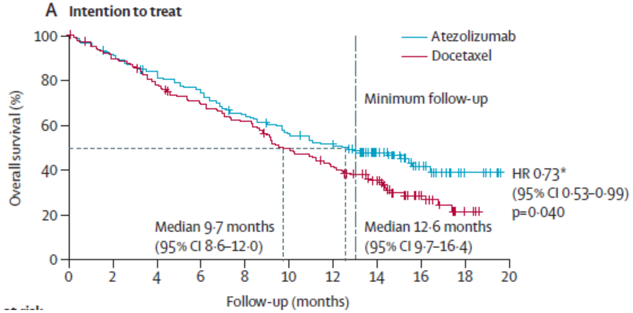
Nivo – renal



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