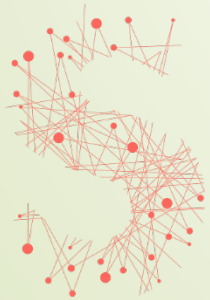
Statistical issues in the development of cancer immunotherapy

Andrew Stone

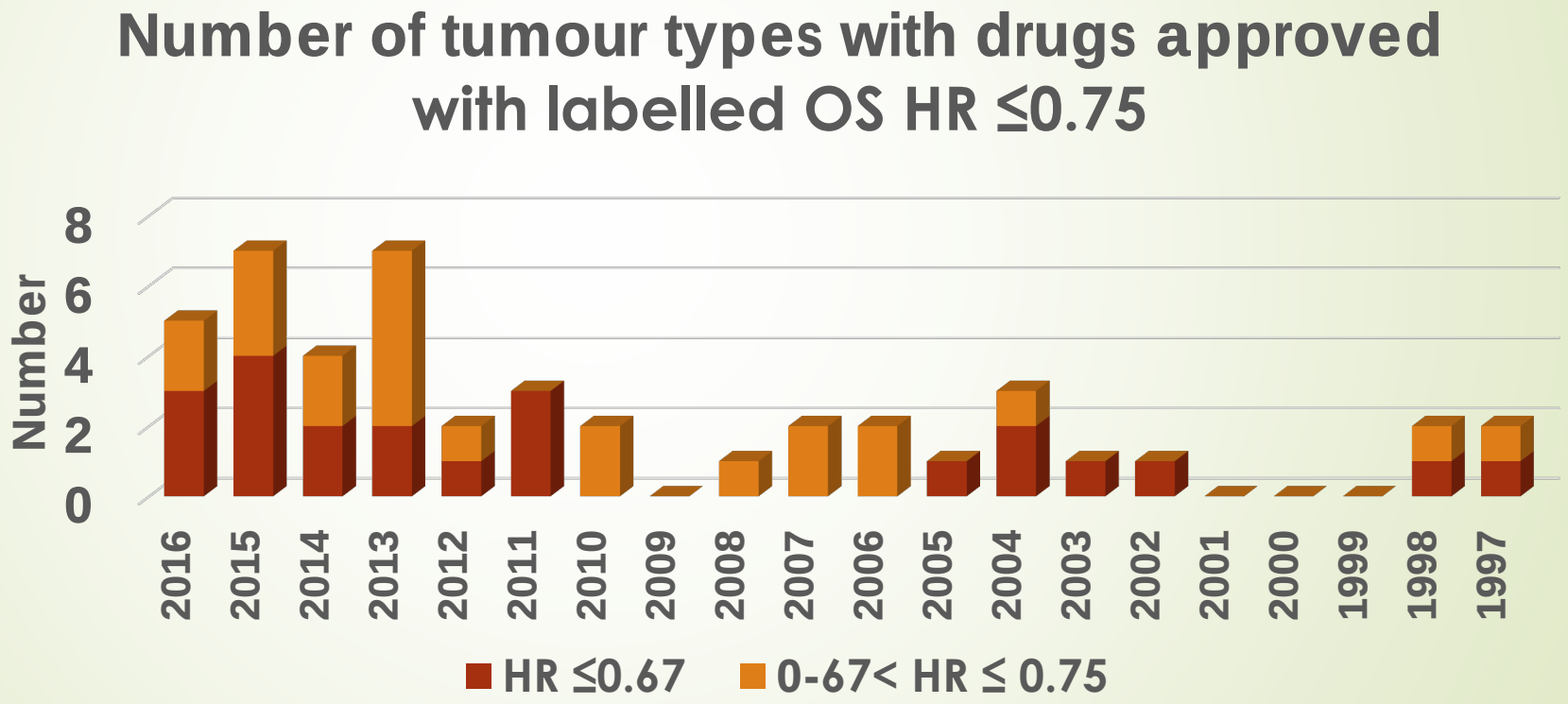
Stone Biostatistics Ltd



stone biostatistics

E: andrew@stonebiostatistics.com
T: +44 (0) 7919 211836
W: www.stonebiostatistics.com

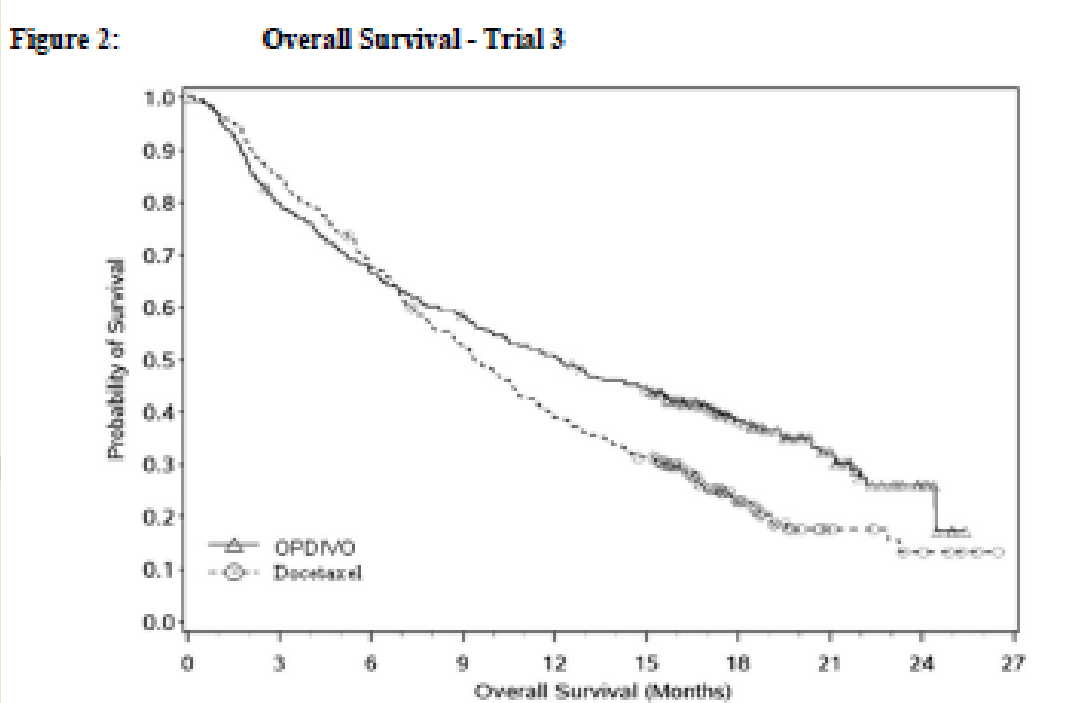
Are we in a golden age for oncology?



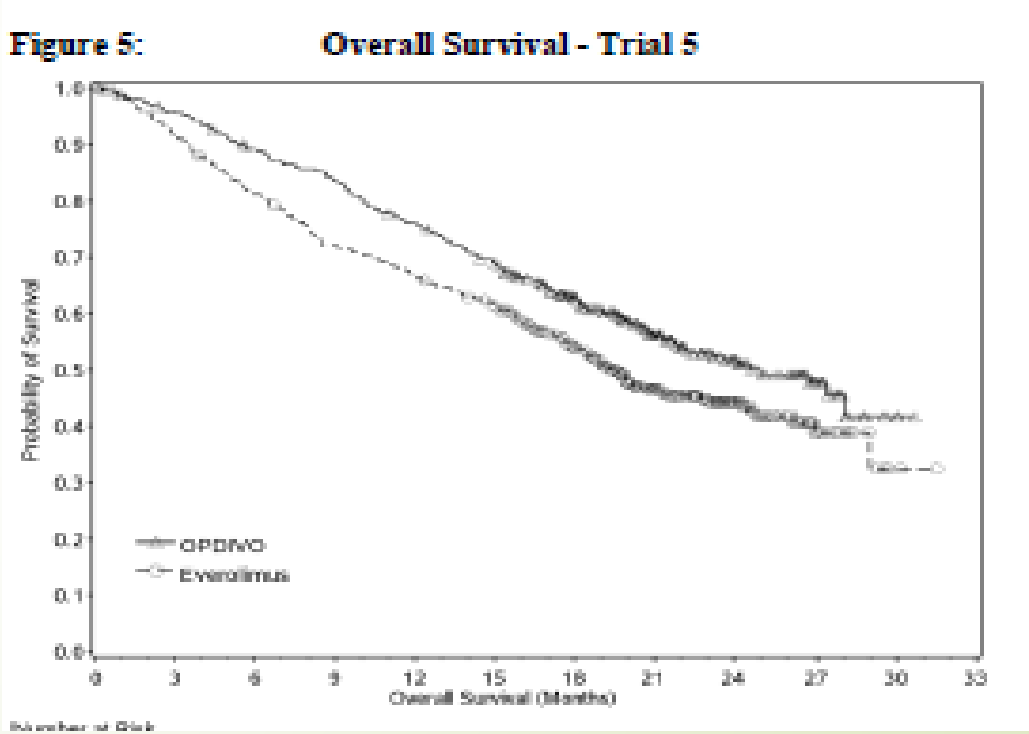
www.stonebiostatistics.com/blog

What's different about IO?

non-proportional hazards (NPH) or more specifically a delayed effect – well, often but not always



Nivolumab 2nd line NSCLC

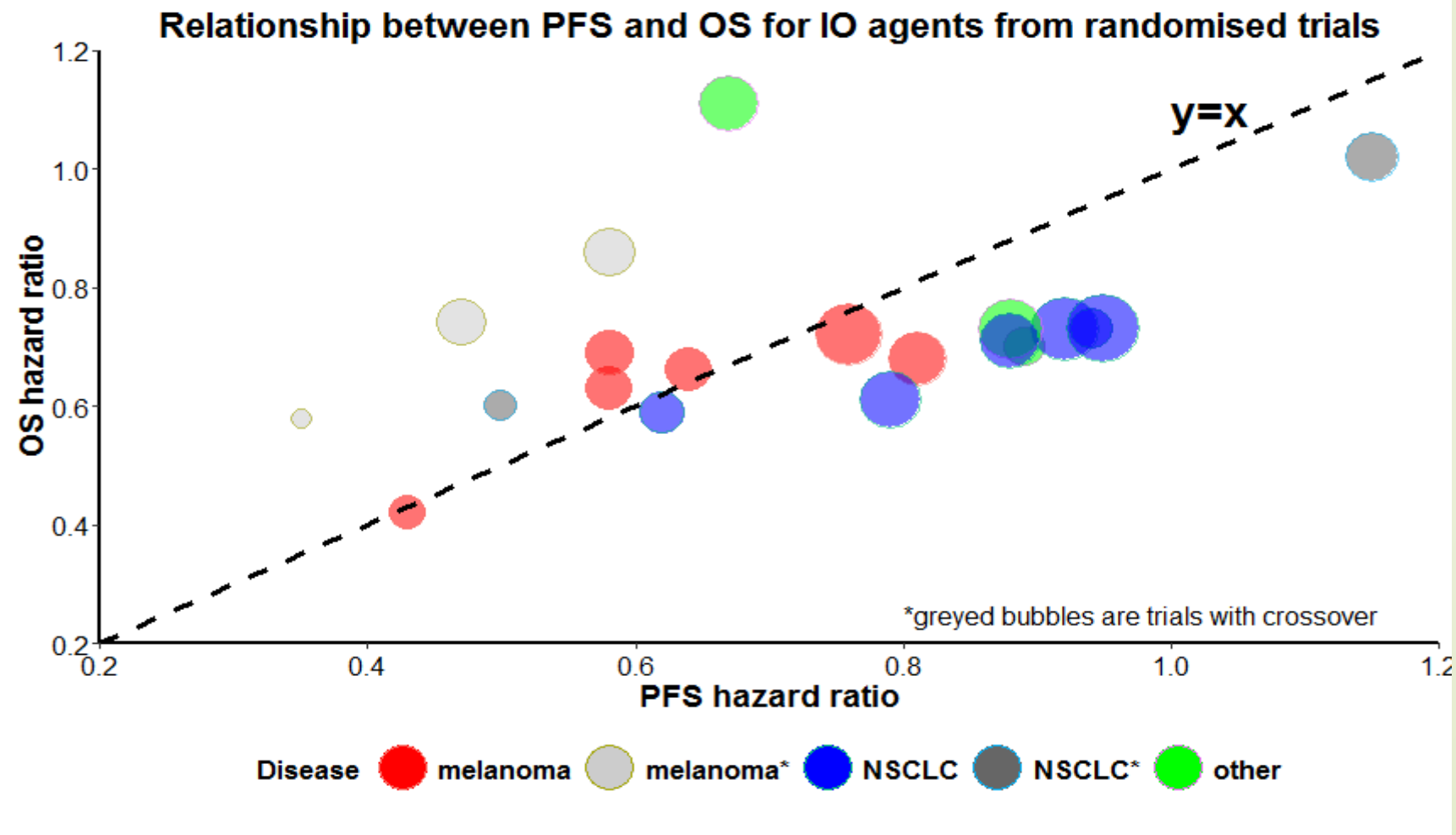


Nivolumab renal

What's different about IO?

Different endpoint relations

Data as of October 2016



Getting the benefit of the doubt?? – bladder cancer

- 5 agents given accelerated approval in the same indication in bladder cancer
- Labelled ORR point estimates range from 13% to 29%
- Labelled lower 95% CI ranges from 9% to 24%
- ORR shown in single arm studies leading to approval is typically higher than this

- Reviews suggest duration of response was the key additional factor in approval
  - FDA comment ‘ *Although the point estimate for the response rate may be lower than what is reported in single-arm studies involving chemotherapy or combination chemotherapy regimens in this disease setting, the durability of the responses observed with XXXX appears to be better than available (off-label) therapy. It is important to note that at the time of this recommendation, the data regarding the durability of response is not yet mature FDA approvals’*

For Discussion

- Analysis of Data
- Sample Size
- Further Design Considerations
- Future Challenges

Analysis of Data

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.

With any type of therapy is it realistic that we ever have truly have PH??

* Uno H, J Clin Oncology 2014 2380-5

HR generated from log-rank or Cox can be interpreted as an average HR

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

$=(U_1+U_2)/(V_1+V_2)$

$= (e_1/4 . \ln(HR_1) + e_2/4.\ln(HR_2)) / (e_1/4 + e_2/4)$

$= p_1 \ln(HR_1) + p2 \ln(HR_2)$

Or overall HR can always be interpreted as the geometric mean of piecewise HRs or more simply:

- the average HR over time, or the average benefit

** Berry G, et al. Statistics in Medicine 1991; 10:749-755.* If HR small say < 0.4, approximation can result in some slight over-estimation (too close to 1) of the true HR

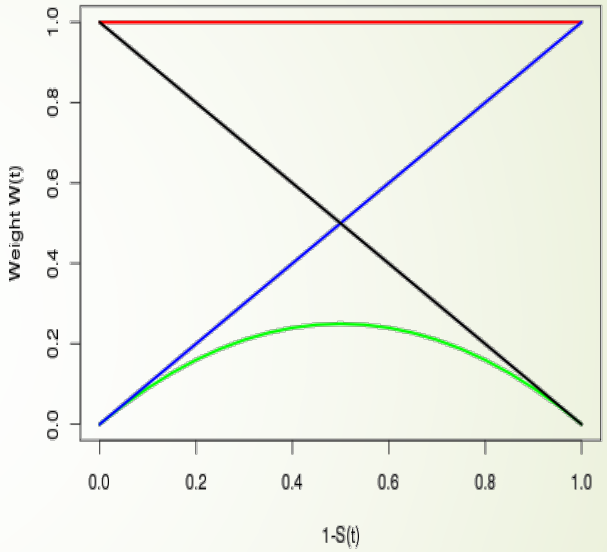
Sellke, T. and Siegmund, D. Biometrika 70: 315-326, 1983

One not so important nuance to be aware of

- Unlike PH, when censoring occurs effects the estimated HR
  - As effects proportion of events observed in each period
  - If only administrative censoring, HR will only be altered if the minimum follow-up, is less than the length of the first period/time-lag
  - Easy to adjust for in sample-size and not a major issue

Yes, but the log-rank is not the most powerful test

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



$$W(T_i) = (\hat{S}(t_i))^{\rho}\ (1 - \hat{S}(t_i))^{t}$$

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

Power can increase with unequal weights assuming you guess correctly

WLRT	Under H_0	No Delay	2-Month Delay	4-Month Delay	6-Month Delay
$G^{0,0}$	4.8	89.9	67.5	43.3	23.5
$G^{1,1}$	4.9	85.9	78.1	55.2	29.8
$G^{0,1}$	5.5	79.4	74.7	60.5	41.2
$G^{1,0}$	5.4	85.8	50.9	24.5	12.1

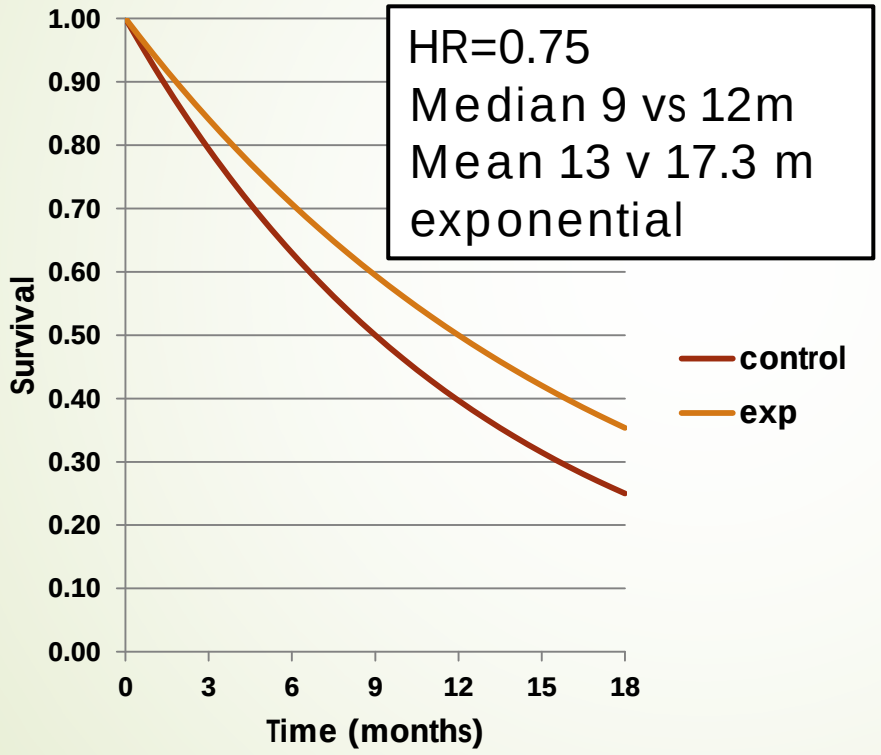
266 patients data analysed after 193 events. True HR=0.625 after displayed delay. 1:1 randomization, 15 month uniform accrual and 7 month control median.

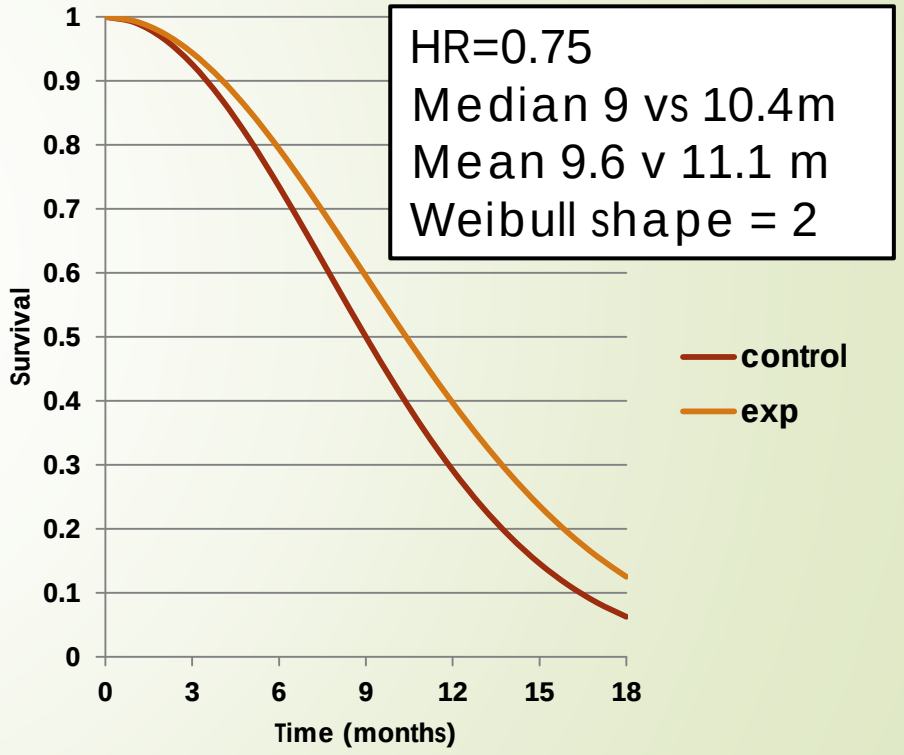
Just because you can – should you?

- If you correctly guess the shape of the curves you can increase power substantially.
- However, are the results clinically meaningful?
  - With WLR(0,1), for example, the weight function would imply it was more important to extend life of the better prognosis patients
  - Why are some patients more important than others
  - So unless you can identify the means to predict which patients will benefit most you'll need to expose the patients with no benefit
  - Is this consistent with an overall benefit/risk??

Measures of benefit & benefit/risk

Even with PH, HRs can correspond to quite different absolute benefit





Restricted Mean

- It was originally proposed to restrict mean to period of PH but no need to
- Decisions on what period
  - Latest event or censored time?
  - Min(latest per arm) etc etc
- Will it be more powerful? – emerging data to say less powerful in situations of delayed effect vs Cox
- Why restrict mean at all?....

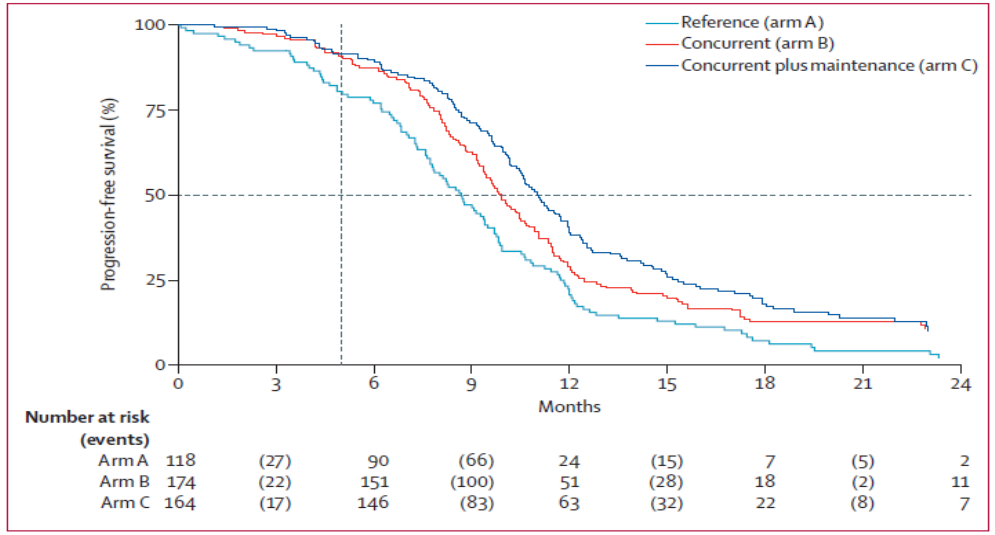


Figure 2: Kaplan-Meier plot of progression-free survival over 2 years
Vertical reference line shows the median time to completion of the chemotherapy phase. Number at risk every 6 months shown with the number of failure events in parentheses, after the time in which the number at risk was calculated.

'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*

Parametric analyses underutilised

- Why not analyse data parametrically and estimate lifetime mean??
  - For example Weibull
  - Mean survival = $e^{\mu}\Gamma(1+\sigma)$, where s is the scale parameter from the model and m the intercept
  - And var(ln(mean)) = var(m) + (digamma(1+s))^{2}var(s) + 2. digamma(1+s)Cov(m,s)
- Unknowns
  - influence of tails and power
- Pretty much every other TA analyses data parametrically
- Important to stress this would be as a supportive measure to the HR to describe absolute benefit

What if patients were cured?

- Could it then be appropriate to approve a therapy if 80% had no benefit over available therapy but 20% were cured?
- Statistical challenges with cure rate models
  - They will always fit and can be badly biased
  - How long should you follow patients, with disease specific survival, to judge there is a cure fraction. Measures available determine how long (piece of string) – some measures available but…
- Simplest approach analyse KM % surviving past x years, where x, based on historical data, is rare
- Analyse as additional primary endpoint, reserving alpha –
  - plan to analyse after OS. May rescue an otherwise –ve trial if a small group of patients have very long survival

Benefit/risk

- What if an agent is equally efficacious as the Standard of Care (SOC) but was better tolerated
- Raises the spectre of non-inferiority
  - This is also present with use of cure rates as well
- Lots of challenges though

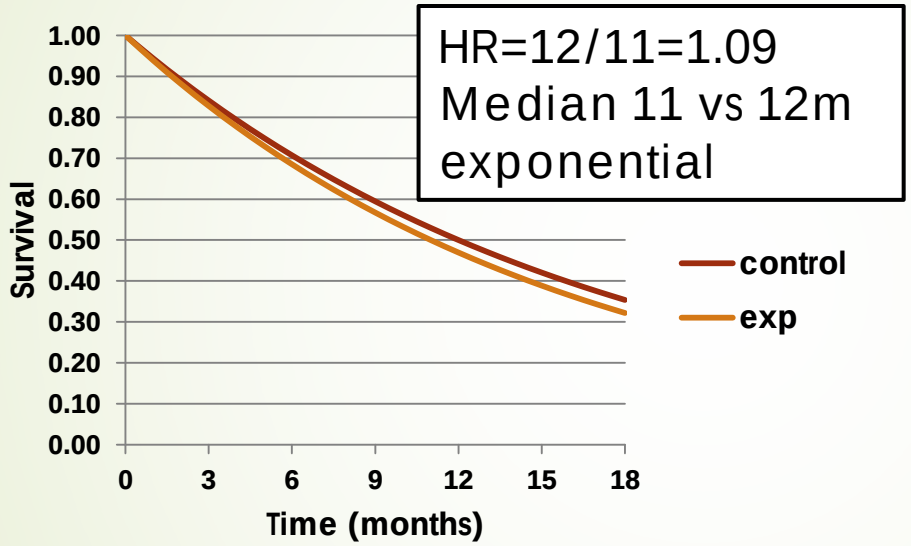
Non-Inferiority when?

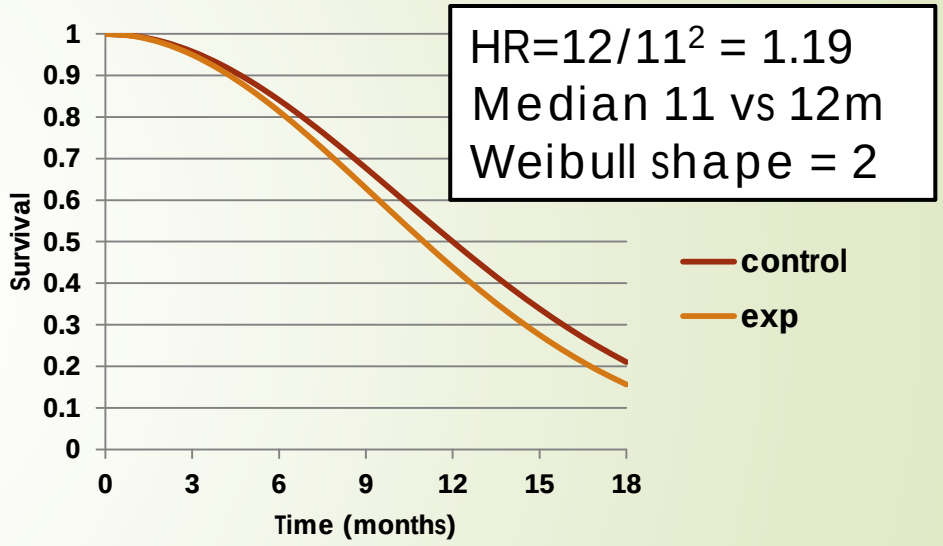
- Situations where:
  - SOC has shown an OS advantage over SOC
    - Doublet and singlet chemos in 1st and 2nd line NSCLC,
    - but not chemo in 1st line melanoma
  - AND can show tolerability benefit on pre-defined, meaningful endpoint
    - Lots of questions here that aren't discussed

NI how?

- Approaches such as effect retention, 95/95 (upper 95% CI less than 95% from meta-analysis of historical trial) opaque to non-statisticians who we need to engage
  - Often required to have shown comparator OS effect in multiple trials
  - This will rarely happen due to ethics of repeating studies
- What about something simple to say – if a drug is better tolerated we would consider approval if we could rule out say
  - 1 month detriment if control group median is < 18 months
  - 2 month detriment otherwise
  - Would then be, as always, a judgement on actual benefit/risk
- There will be, and should be huge debate on such margins

Defining a limit, depends on the shape of curves (or variability in survival)





If observe HR <0.91 upper 95% CI <1.09, with 400 events

If observe HR <0.97 upper 95% CI <1.19, with 400 events

Yes but

- How do you know what shape to propose
- What if it differs in the actual trial
- So why not just use an exponential distribution as a worst case (assuming event rate does not reduce over time with comparator)
- Or possibly calculate lifetime-means and construct CI for difference between them

Lots of complications but is it worth it?

- Current model seems to make it extremely difficult to replace poorly tolerated therapies with equally effective but better tolerated ones
  - The hurdles are higher in diseases with the most unmet need
- Do patients deserve a choice?
- There are lots of complications but should we try to tackle as a community
- Currently patients with most need of therapies we're making it harder to provide alternatives
  - Benefits might extend beyond clinical and into the economic with increased competition

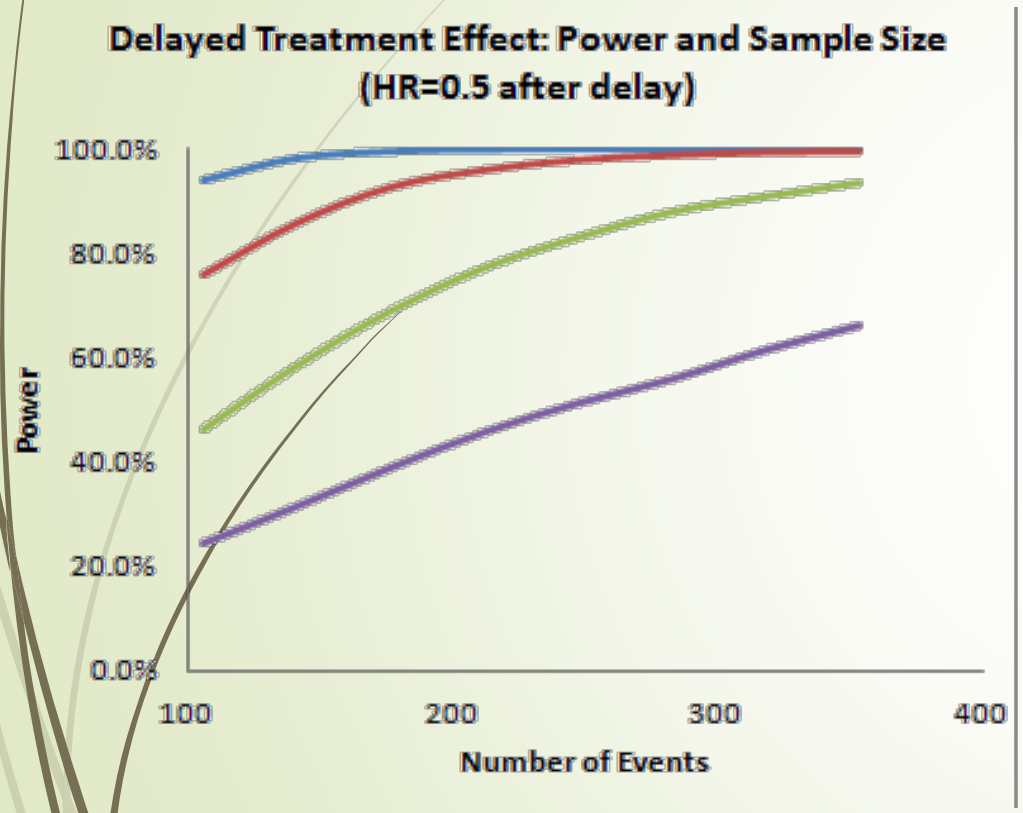
Sample-Size

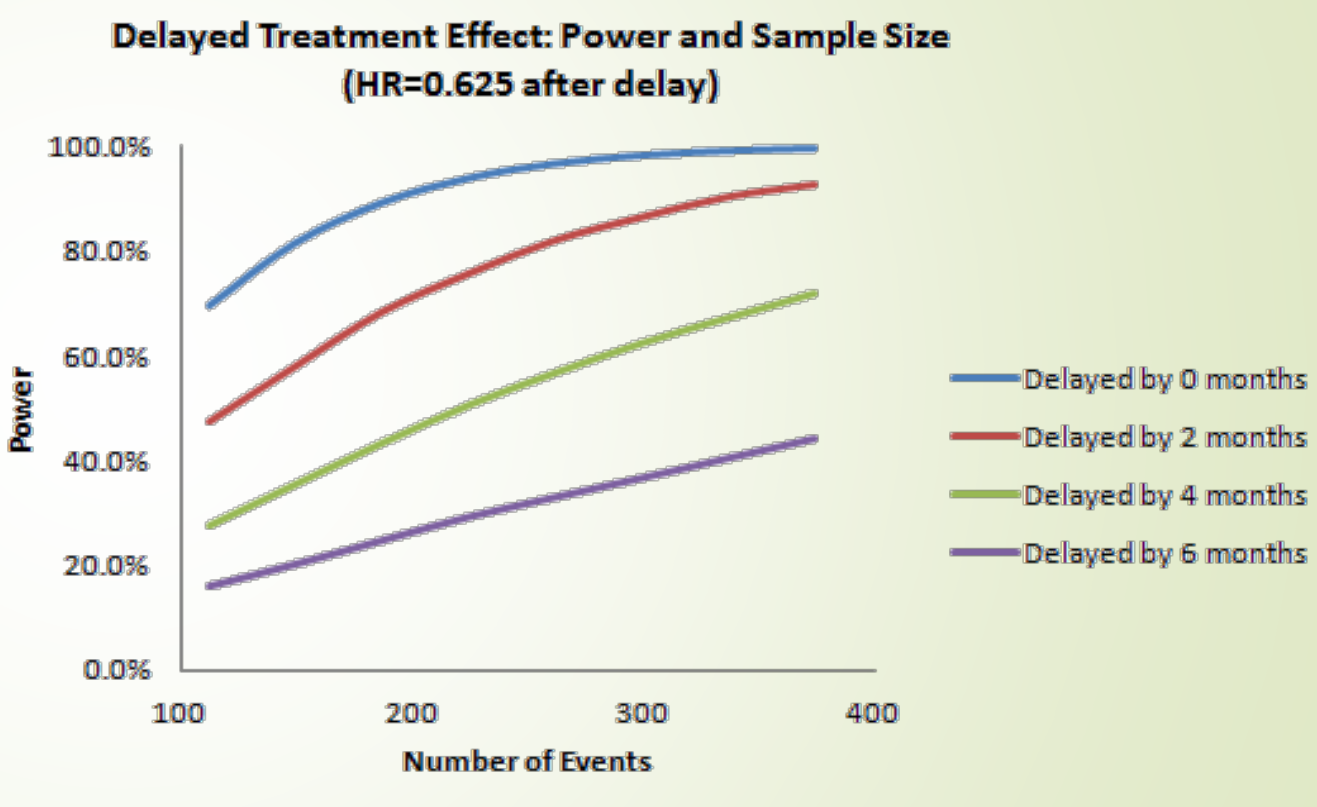
Sample Size with NPH

- For a given follow-up, lag-time, piecewise HRs and recruitment rate calculate:
  - the average HR
  - The expected number of events observed
- The power for this analysis matches the power for PH analysis with given number of events and HR= average HR
- More exactly
  - $If\ S(t) = \begin{cases} \exp(-\lambda_1 t^\gamma) & t < T \\ \exp(-\lambda_1 T^\gamma - \lambda_2 (t^\gamma - T^\gamma)) & t \geq T \end{cases}$
  - $Proportion\ recruited, non-uniformly, by\ time\ s, governed\ by\ k\ G(s) = \frac{s^k}{B^k}$
  - $p(event\ by\ time\ t) == \left(\frac{\min(t,B)}{B}\right)^k - \frac{k}{B^k}\int_0^{\min(t,B)} s^{k-1} S(t-s)\, ds$

More generally with non-uniform accrual: Carroll KJ (2009), *Pharmaceutical Statistics, 8, 333–345. A closed form solution is presented with T=0, exponential and integer k*

Adverse impact on power if delay is not accounted for

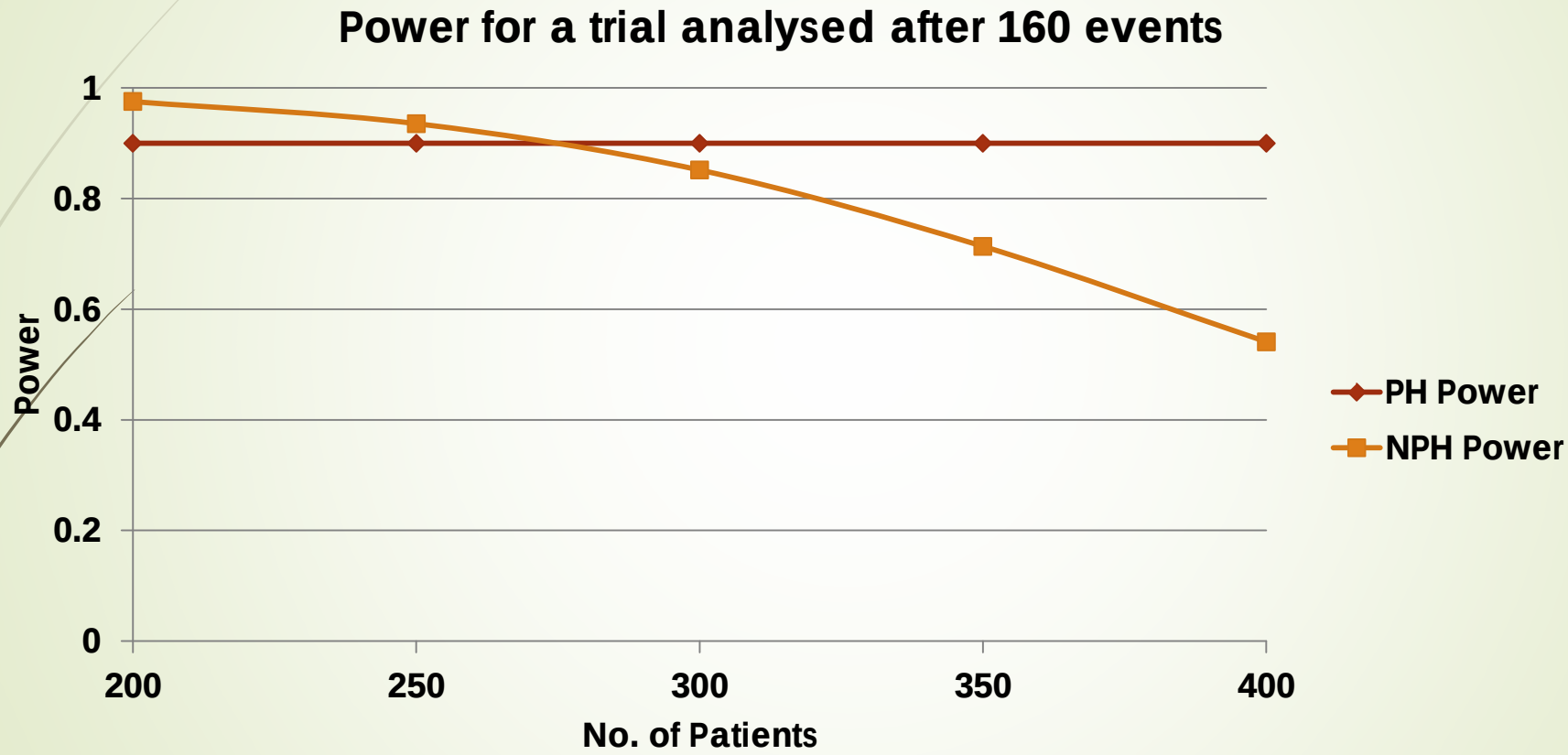




1:1 randomization; assumes e/0.71 patients are recruited where e = no. of events
Fixed 15 months accrual time (uniform);
Median OS (control)=7 months; 2-sided type I error =0.05;

Unlike PH, power dependent on maturity

With NPH, power increases with proportion of events after time-lag



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)**
- An alternative (Fredrik Ohrn AZ) is to re-weight events before and after the lag at the futility analysis to match the expected split at the final analysis
  - This could minimise the delay resulting from 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

Further Design Considerations

Contribution of Components (CoC)

- Increasingly two unapproved therapies are being investigated, eg combination of IO therapies
- In addition to showing that the combination regimen has better B/R than comparator, need to demonstrate both agents are needed
- How should this be done?

How

- We need to establish each agent is an active contributor to the overall efficacy of the combination (A+B)?
- What should this entail?
  - If the primary endpoint for comparing combo(A+B) to SOC is OS, it should not mean independently showing A+B is better to both A and B on OS
  - If that was shown on PFS or RR that should be sufficient
  - Could it even be shown on average tumor size or functional PD endpoint such as DCE-MRI/PET etc?
- Note: may also not be appropriate to randomise if one of constituents has shown a very low RR in a closely related indication or consider possibility of dropping that arm early

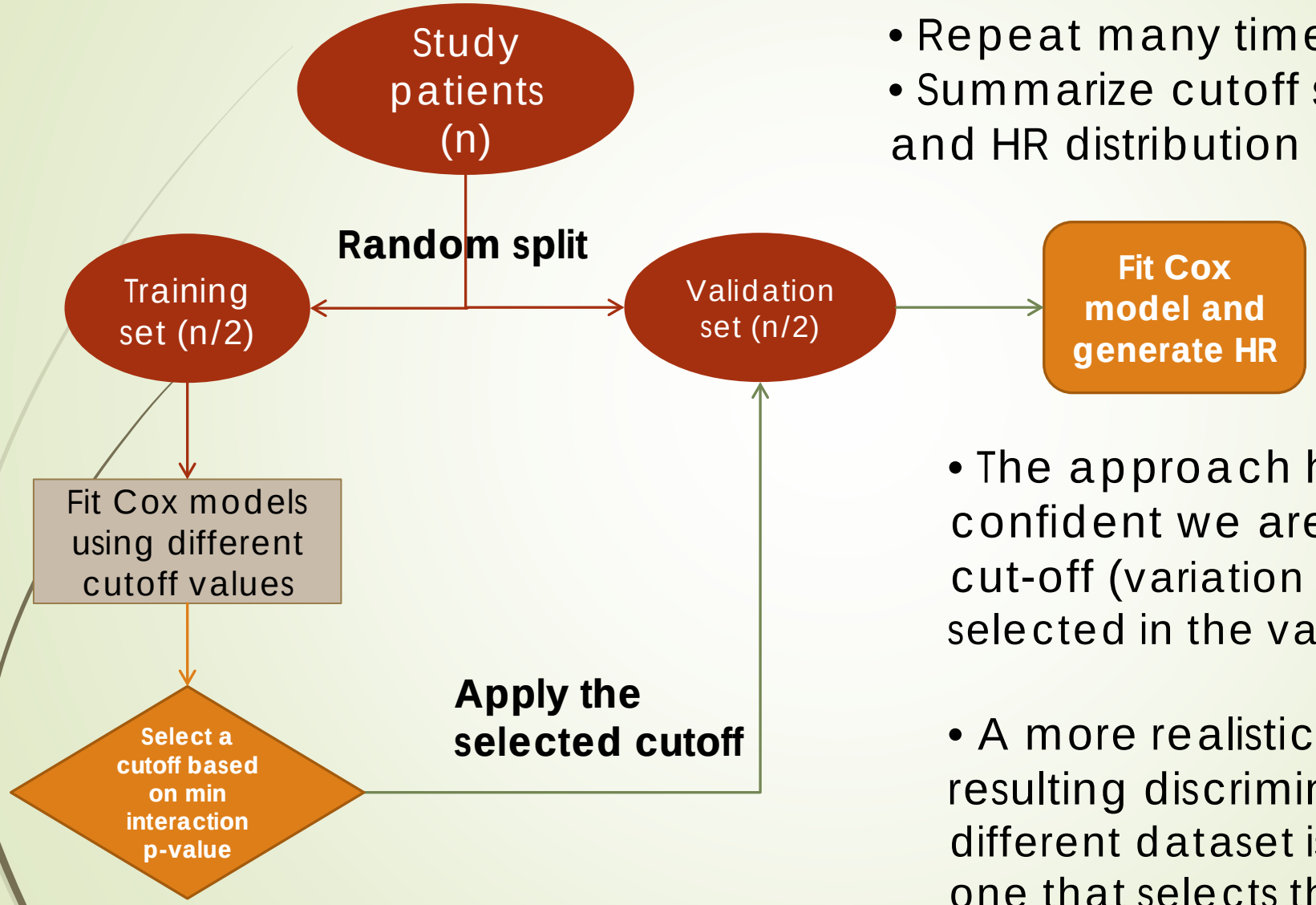
Subpopulations

- As with most agents the effect of therapy may differ according to patient subgroup eg PDL1 expression
- If add a primary endpoint, small price to pay on sample size to maintain power on the original endpoint
- As only need to increase sample size by **7%** for a **4%/1%** split
  - A good strategy if I have a biomarker defined subgroup, gets 1%, which would likely have a bigger treatment effect anyway ie +ve trial if drug only works in a subgroup
- And **21%** increase in SS for **2.5%/2.5%** split (ie not double as sometimes assumed)

- I often encourage hedging – little price to pay in terms of sample-size and that decision might rescue a negative trial (if that endpoint/group is approvable)
  - Including hedging between overall and sub-population in primary analysis and within hierarchy of secondary endpoints – we're often surprised

*Assumes 80% power, increases slightly less with 90% power

Biomarker Cutoff Optimization – cross validation



- Repeat many times
- Summarize cutoff selection and HR distribution

- The approach highlights how confident we are of the correct cut-off (variation on the cut-off selected in the validation set)

- A more realistic view of the resulting discrimination (as a different dataset is used to the one that selects the best cut-off)

Multiplicity

- Testing strategies rapidly get very complicated
- Critical for statisticians to work through this carefully with clinical, regulatory and commercial colleagues
- Hedge bets in hierarchy
- Consider when need to adjust – testing CoC may not need to be in hierarchy

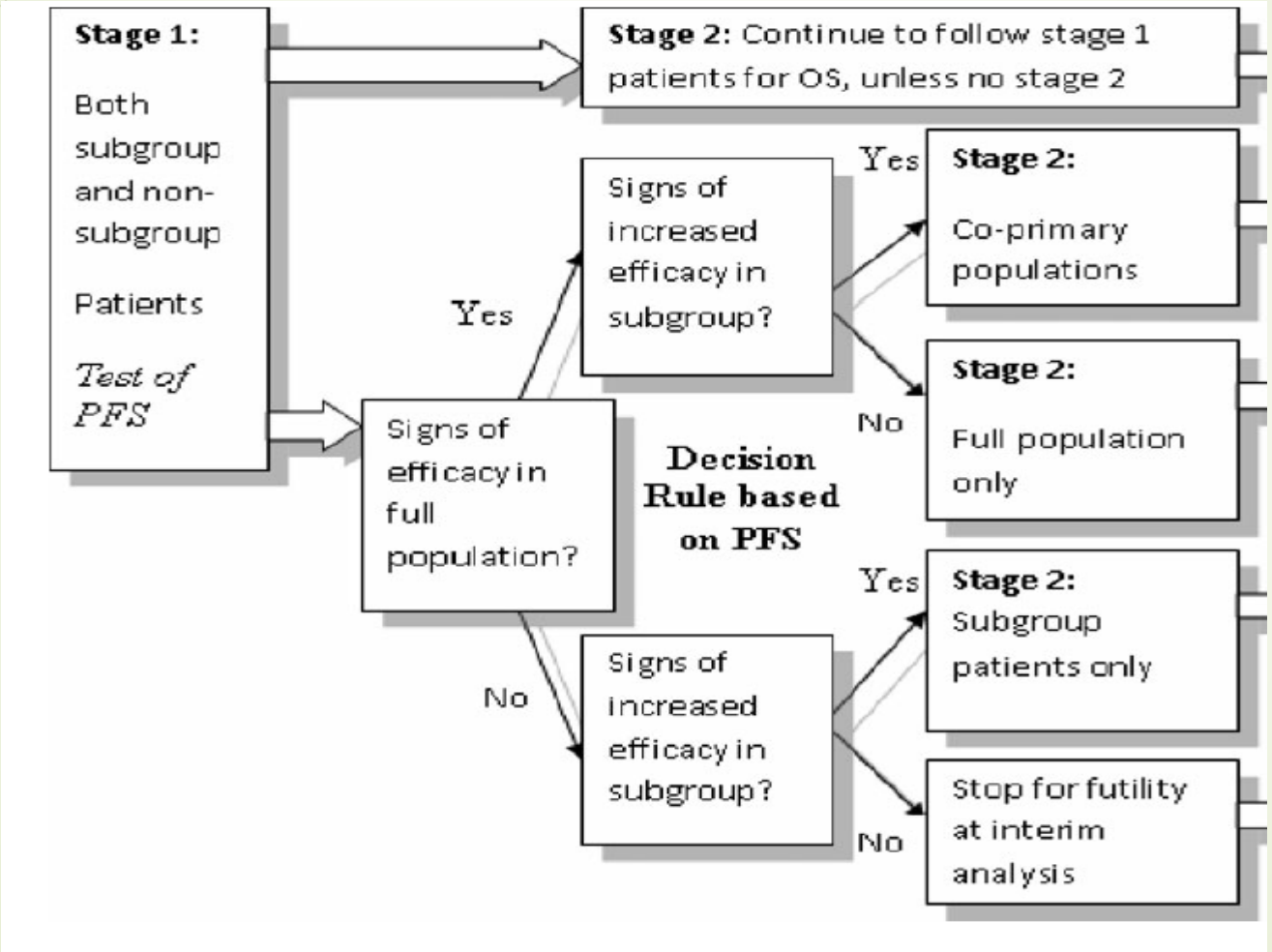
An alternative approach to using single-arm trials as a first step

- Checkmate 037: Randomised (2:1) comparison of nivolumab vs SOC in ≥2nd line melanoma
- Planned interim, non-comparative analysis of ORR in nivolumab in first 120 patients with ≥6 month F/U
  - Demonstrated ORR = 31.7% (23.5%, 40.8%)
- Led to accelerated approval in US by itself, and contributed to EU approval with results of completed randomised trial in 1st line
  - Although, interim OS from this trial, presented in EPA, HR=0.93 (0.68, 1.26)

- A better approach than relying on single-arm trials, especially as commitment trial will be well under way and complete soon after

http://www.ema.europa.eu/docs/en_GB/document_library/EPAR_-_Public_assessment_report/human/003985/WC500189767.pdf

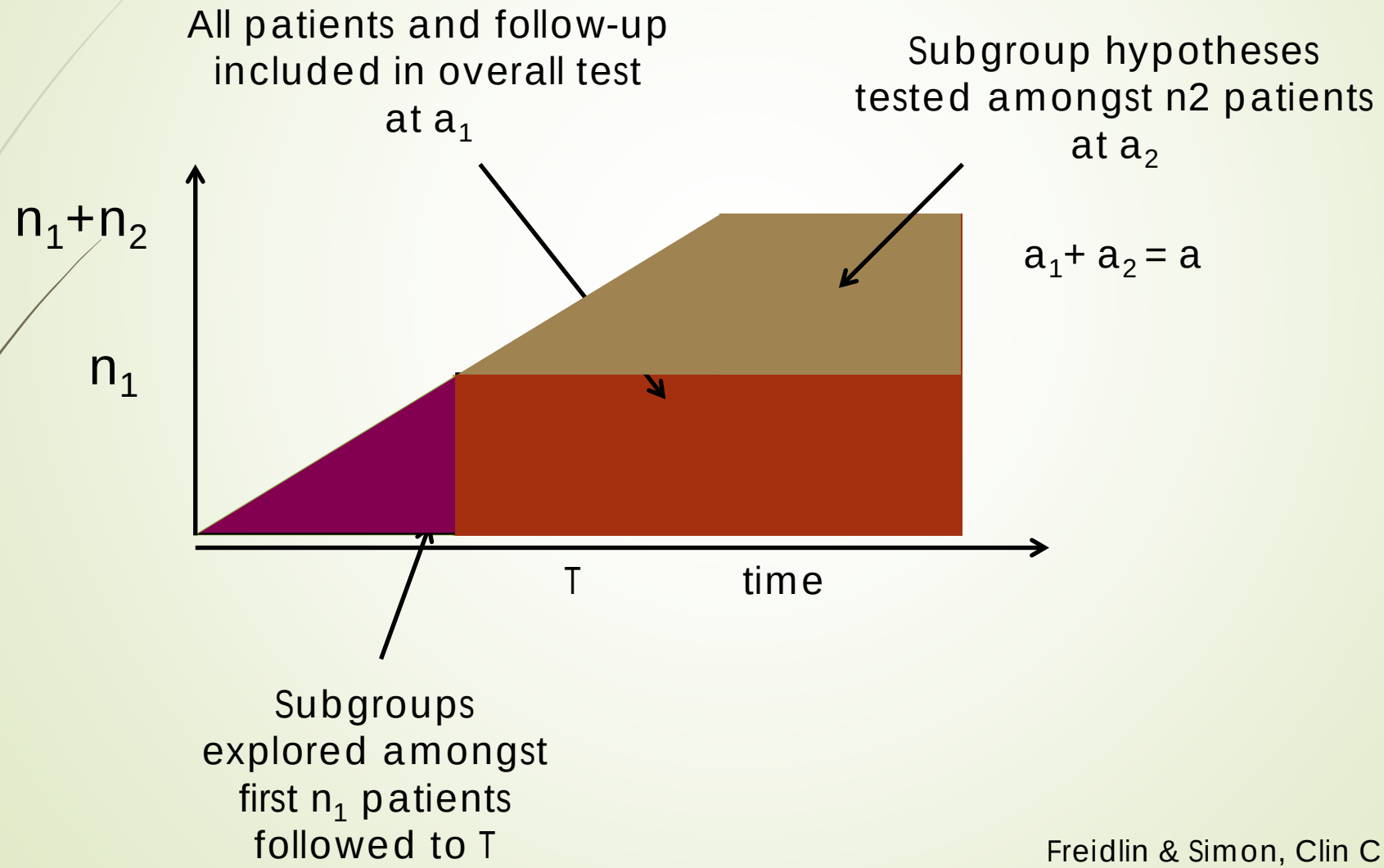
Adaptive sub-population design: some idea on population but want flexibility



Jenkins M, Stone A, Jennison C. *Pharm. Statistics 2011 4 347-356*

Adaptive signature design – an underutilised approach

Learn and confirm within the same design



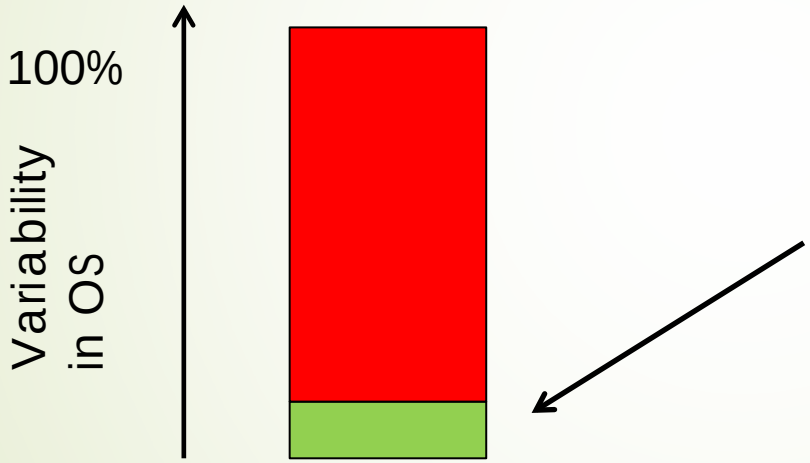
Freidlin & Simon, Clin Cancer Res 2005;11(21) 2005

Challenges on the Horizon

- Indications for IO will run out
- Later entrants will suffer from cross-over to other IO
  - May block entrants in indications reliant on OS and when PFS does not predict OS
  - Likely face situations where PFS effect replicates other IOs but OS blurred/not shown by cross-over to other IO in control arm
    - Then what do?

Will analyses that adjust for OS have a role?

- IPCW and RPSFT make very strong assumptions – amount of variability explained by independently prognostic covariates is very low



- Control group of vandetanib vs Best Supportive care in NSCLC[#]
- % of variability explained by :

Tumour stage (IIIb, IV),
No. of organs involved (1 or 2, 3 or more),
Histology (adenocarcinoma, squamous, other),
WHO performance status (0 or 1, 2 or higher),
Smoking history (smoker, non-smoker),
Gender (male, female),
Race (Caucasian, oriental, other),
Prior EGFR-TKI (erlotinib, getinib),
EGFR expression (+, -, unknown) ,
EGFR gene amplication (FISH) (+, -, unknown)
EGFR mutation status (+, -,unknown).

- factors highlighted independently prognostic

[#] *JCO April 1, 2012 vol. 30 no. 10 1114-1121 Analysis presented at AAADV conference May 2015*

Randomisation preserving OS adjusted analyses – a (better?) alternative

- If the experimental agent extended OS, then the treatment effect in centres/countries with least or absent cross-over should show the biggest treatment effect
- Therefore, look at OS result, at centre level, by degree of crossover (present/absent, by quartiles of centre use etc)
- Key is that expect unmeasured confounders balances at centre level on average as randomisation preserved (Kaiser*)
- Could also include all centres but censor all patients on first use of subsequent IO in that centre
  - Eg first use of subsequent IO on July 1st 20XX, censor any patient who has yet to die at July 1st 20XX, excluding patients randomised after that

Pharmaceut. Statist.* ***2013,* 12 43–47

IO challenging conventions

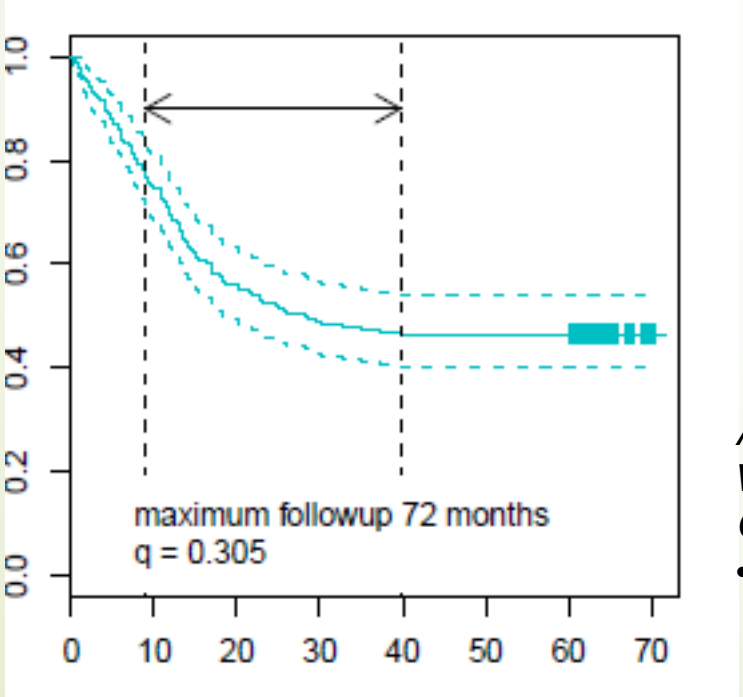
- √ How we size
- √ How quantify benefit
- √ Cure rates
- √ NI
- √ More sophisticated designs

- × But not our primary analysis method

Back-ups

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 observat
Whereas if the denominator was the number of uncensored
Observations would have better properties
- *max value =1 independent of cure rate*