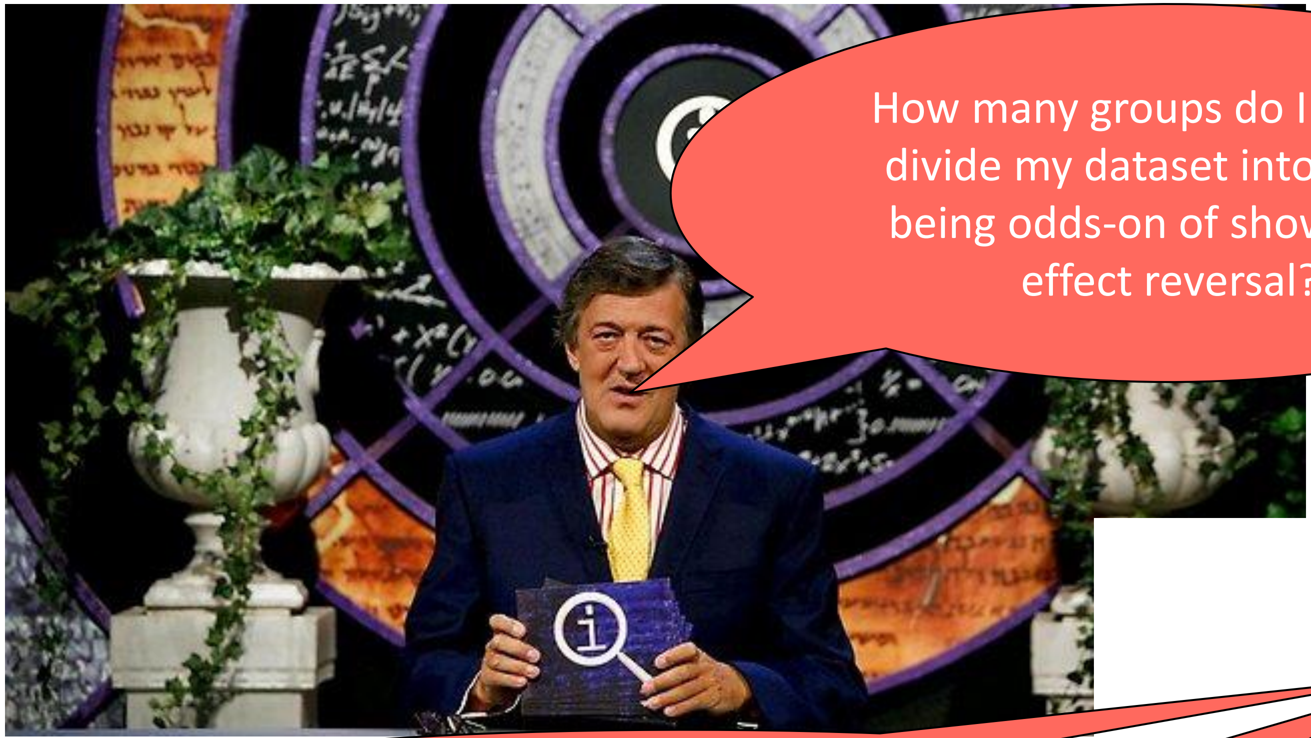
Predictive Biomarkers in Drug Development

Andrew Stone

StoneBiostatistics Ltd.

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

Statistical QI



How many groups do I need to divide my dataset into before being odds-on of showing an effect reversal?

Only 6

If the treatment effect is identical per group and equal to the hypothesised treatment effect H1. Oh, and assuming equal n per group & trial has 80% power

Senn, Statistical Issues in drug development

Statistical QI

If instead of powering for a 10 unit treatment effect , I powered for 10 unit difference in treatment effect between males and females, how much bigger is the trial?

4

Assuming 50% were male

Byar, SIM, 1985, 255-63

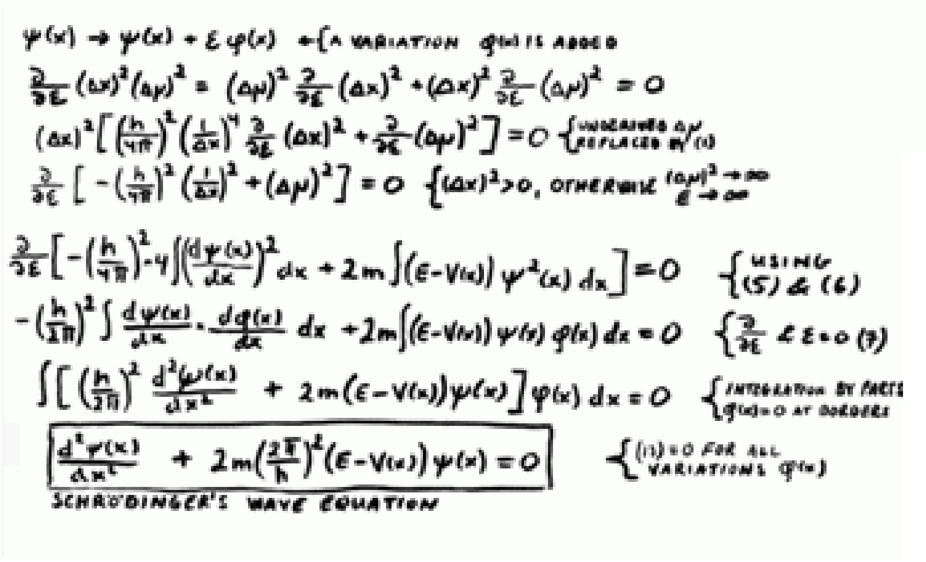
Subgroups – what we can be certain of!

I told you so!

Oh I didn’t tell anyone beforehand

We have a serious false+ve problem

And we have a serious false-ve problem



$\psi(x) \rightarrow \psi(x) + \varepsilon\varphi(x)$ ← {A VARIATION $\varphi(x)$ IS ADDED

$\frac{\partial}{\partial\varepsilon}(\Delta x)^2(\Delta p)^2 = (\Delta p)^2\frac{\partial}{\partial\varepsilon}(\Delta x)^2 + (\Delta x)^2\frac{\partial}{\partial\varepsilon}(\Delta p)^2 = 0$

$(\Delta x)^2\left[\left(\frac{h}{4\pi}\right)^2\left(\frac{1}{\Delta x}\right)^4\frac{\partial}{\partial\varepsilon}(\Delta x)^2 + \frac{\partial}{\partial\varepsilon}(\Delta p)^2\right] = 0$ {UNCERTAINTY Δp REPLACED BY (1)

$\frac{\partial}{\partial\varepsilon}\left[-\left(\frac{h}{4\pi}\right)^2\left(\frac{1}{\Delta x}\right)^2 + (\Delta p)^2\right] = 0$ {$(\Delta x)^2 > 0$, OTHERWISE $(\Delta p)^2 \rightarrow \infty$, $E \rightarrow \infty$

$\frac{\partial}{\partial\varepsilon}\left[-\left(\frac{h}{4\pi}\right)^2 4\int\left(\frac{d\psi(x)}{dx}\right)^2 dx + 2m\int(E-V(x))\,\psi^2(x)\,dx\right] = 0$ {USING (5) & (6)

$-\left(\frac{h}{2\pi}\right)^2\int\frac{d\psi(x)}{dx}\cdot\frac{d\varphi(x)}{dx}\,dx + 2m\int(E-V(x))\,\psi(x)\,\varphi(x)\,dx = 0$ {$\frac{\partial}{\partial\varepsilon}$ & $\varepsilon = 0$ (7)

$\int\left[\left(\frac{h}{2\pi}\right)^2\frac{d^2\psi(x)}{dx^2} + 2m(E-V(x))\psi(x)\right]\varphi(x)\,dx = 0$ {INTEGRATION BY PARTS $\varphi(x)=0$ AT BORDERS

$$\boxed{\frac{d^2\psi(x)}{dx^2} + 2m\left(\frac{2\pi}{h}\right)^2(E-V(x))\,\psi(x) = 0}$$ {(11) = 0 FOR ALL VARIATIONS $\varphi(x)$

SCHRÖDINGER'S WAVE EQUATION

But why would all patients have the same benefit and risk ??

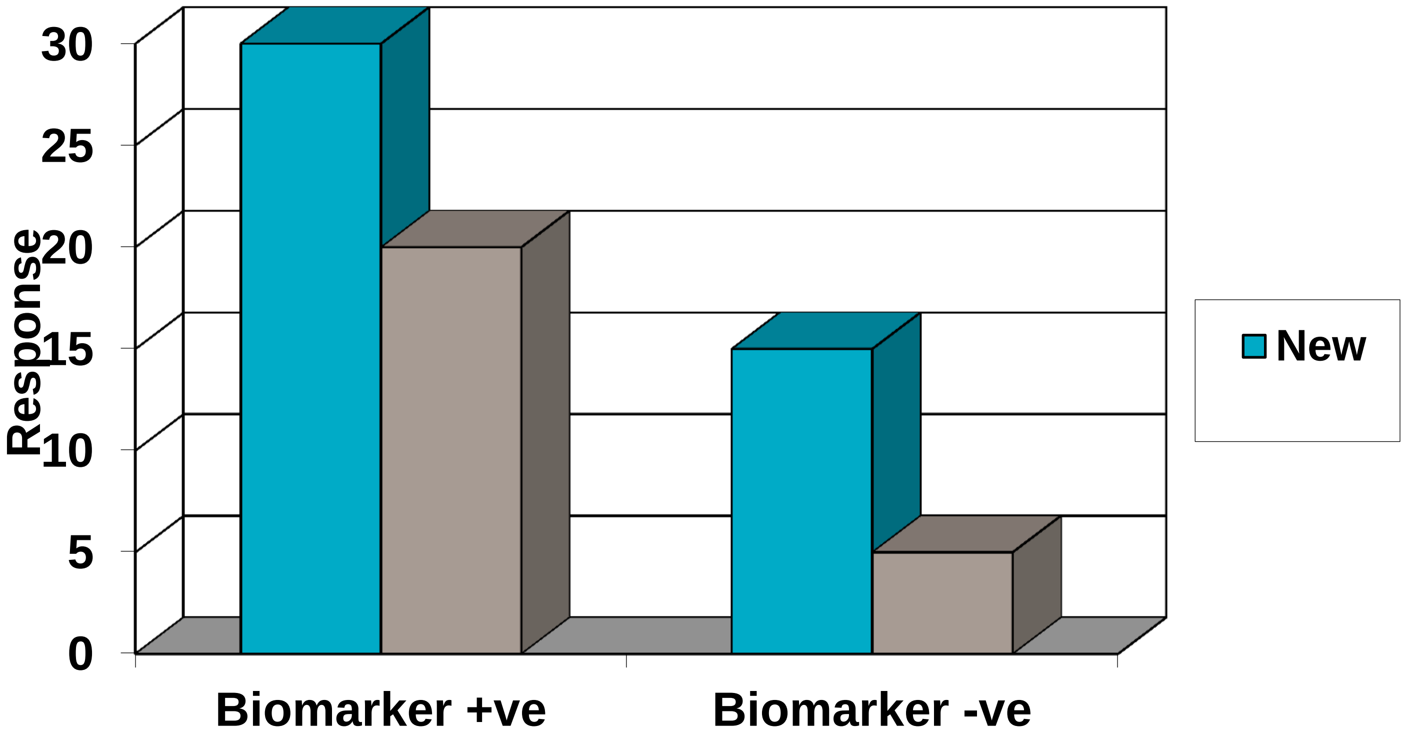
Outline

Overview of

- How best to learn
- How best to confirm
- Learn and confirm
- Case study
- Very briefly, post-treatment biomarkers

How best to learn

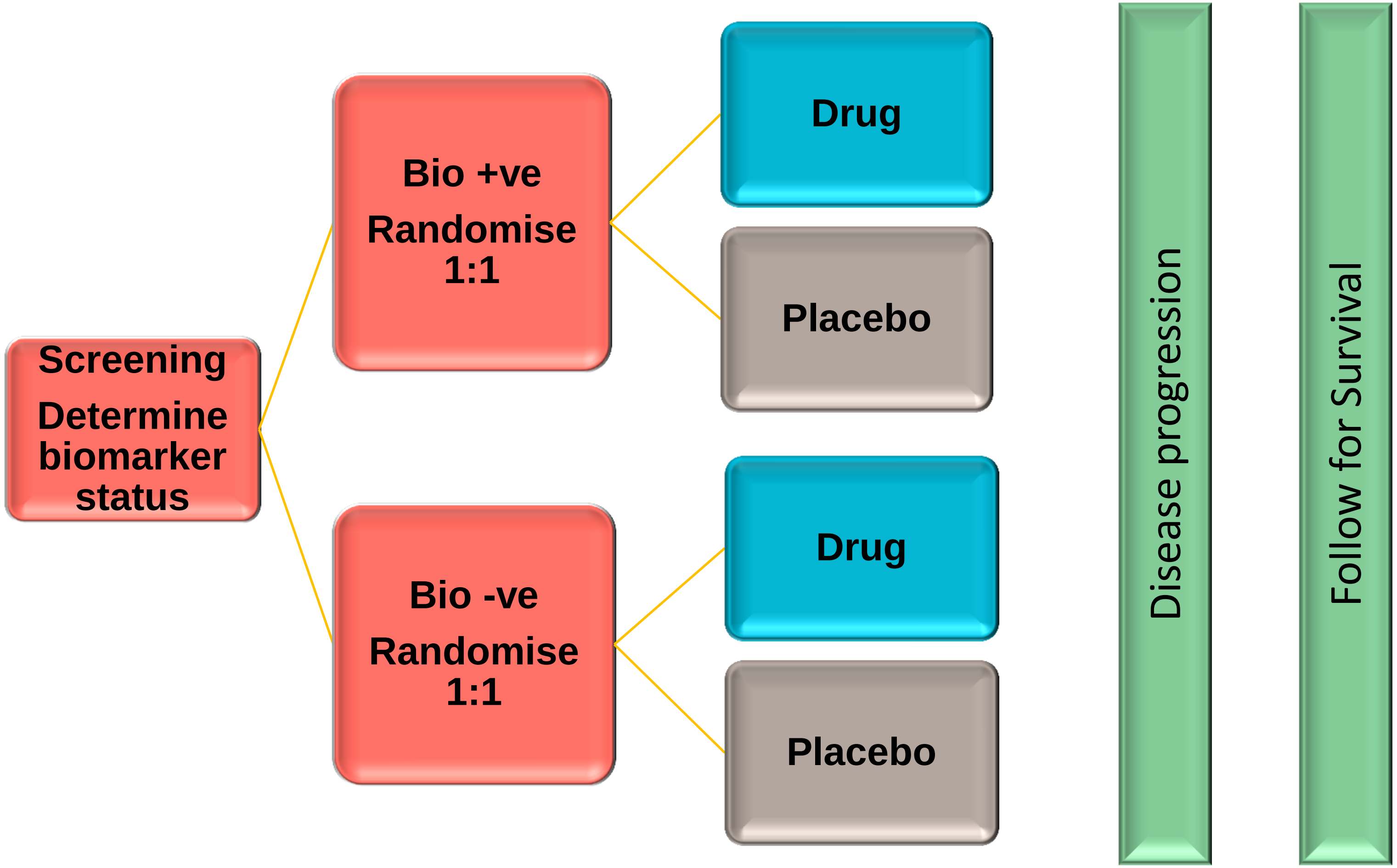
A common oversight amongst colleagues



- Often see comparisons of efficacy within arm,
  - either single arm or
  - within randomised studies
- purporting to show that drug X works better in Biomarker Y

Prognostic Biomarker identified

Better to randomize



Screening

Determine biomarker status

Bio +ve

Randomise 1:1

Drug

Placebo

Bio -ve

Randomise 1:1

Drug

Placebo

Disease progression

Follow for Survival

POPLAR trial with atezolizumab an excellent example.

Findings likely to be exaggerated

- Pre-specify how likely there is to be a treatment-by-subgroup interaction of given size

- Use a Bayesian approach to quantify likely over-estimation of effect*

- Extends following basic result for normal data
  - Prior ~ N(μ,τ^2) , Likelihood ~ N(δ,σ^2) - in this case μ = 0 and τ^2 is pre-specified.
  - *Posterior*
    - mean = $(\sigma^2/\{\sigma^2+\tau^2\}).\mu + (\tau^2/\{\tau^2+\sigma^2\}).\ \delta$
    - Variance = $\tau^2\sigma^2/\{\sigma^2+\tau^2\}$

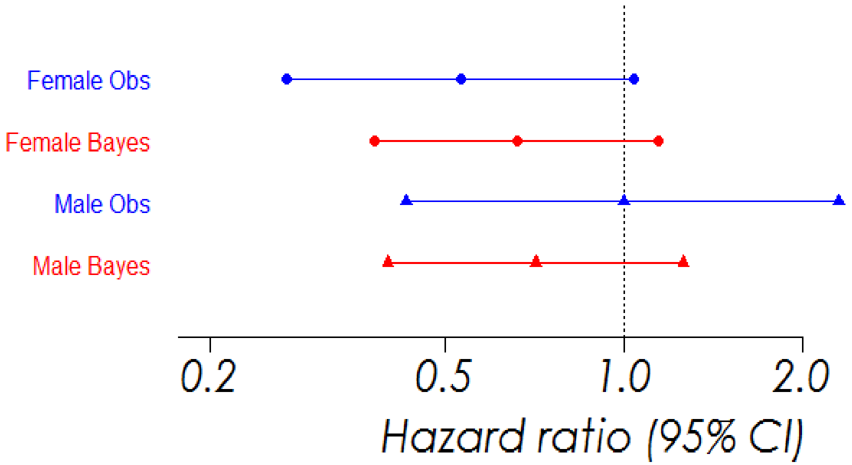
*Simon, SIM 2002 2909-16.

Related non-Bayesian approach by Senn in Chapter 9 Statistical Issues in Drug Development

A simple example:
borrowing information from other subgroups

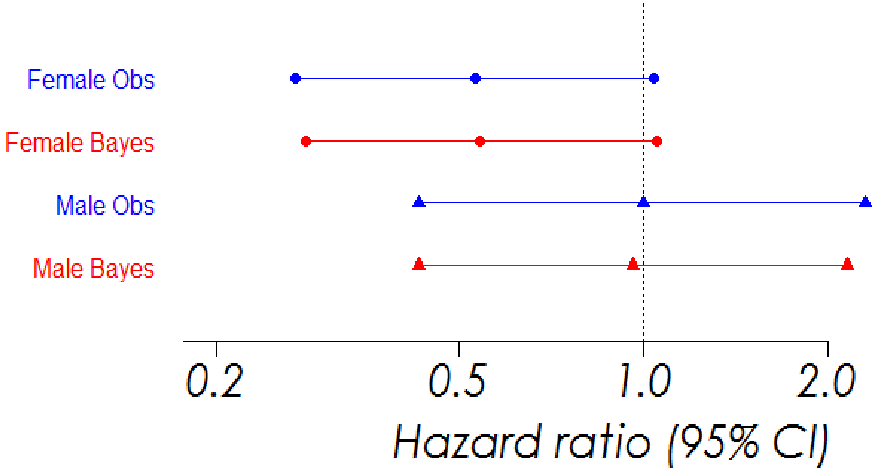
2.5% probability of interaction*

Female bayes = 0.67Female + 0.33male



60% probability of interaction*

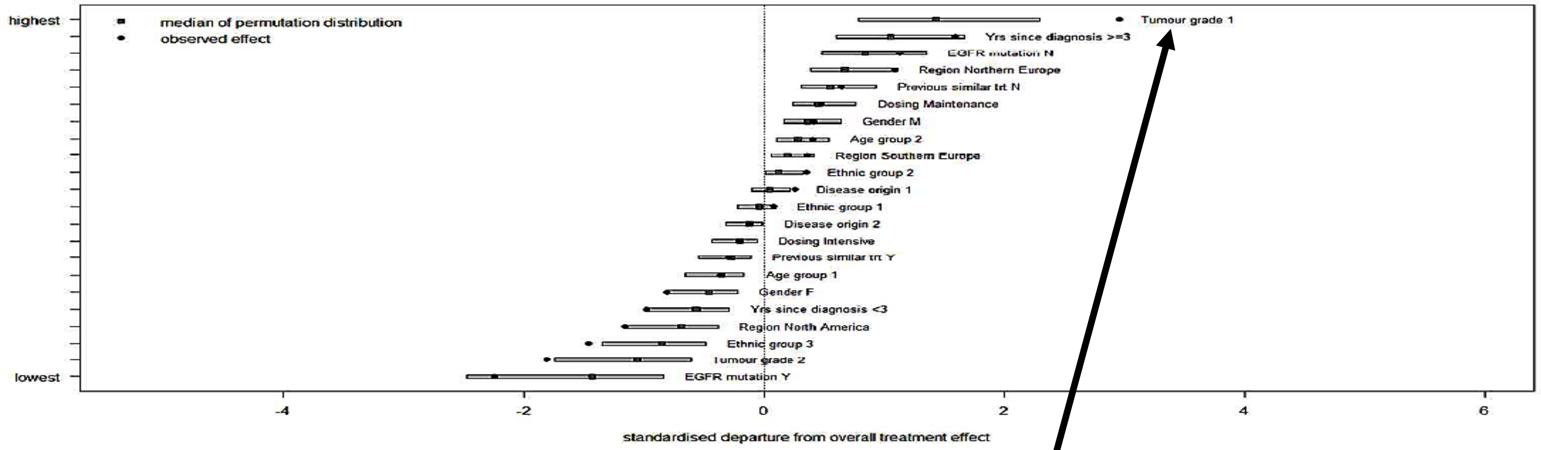
Female bayes = 0.96Female + 0.04male



* Probability HRfem/HRmale =0.67

- Borrow information from other subgroups based on the extent to which an interaction was predicted
  - Consistent with how we interpret such data
- Do not interpret literally but use to give a guide of likely effect shrinkage
  - Based on conversation had prior to unblinding
- More sophisticated methods to cover >1 subgroup

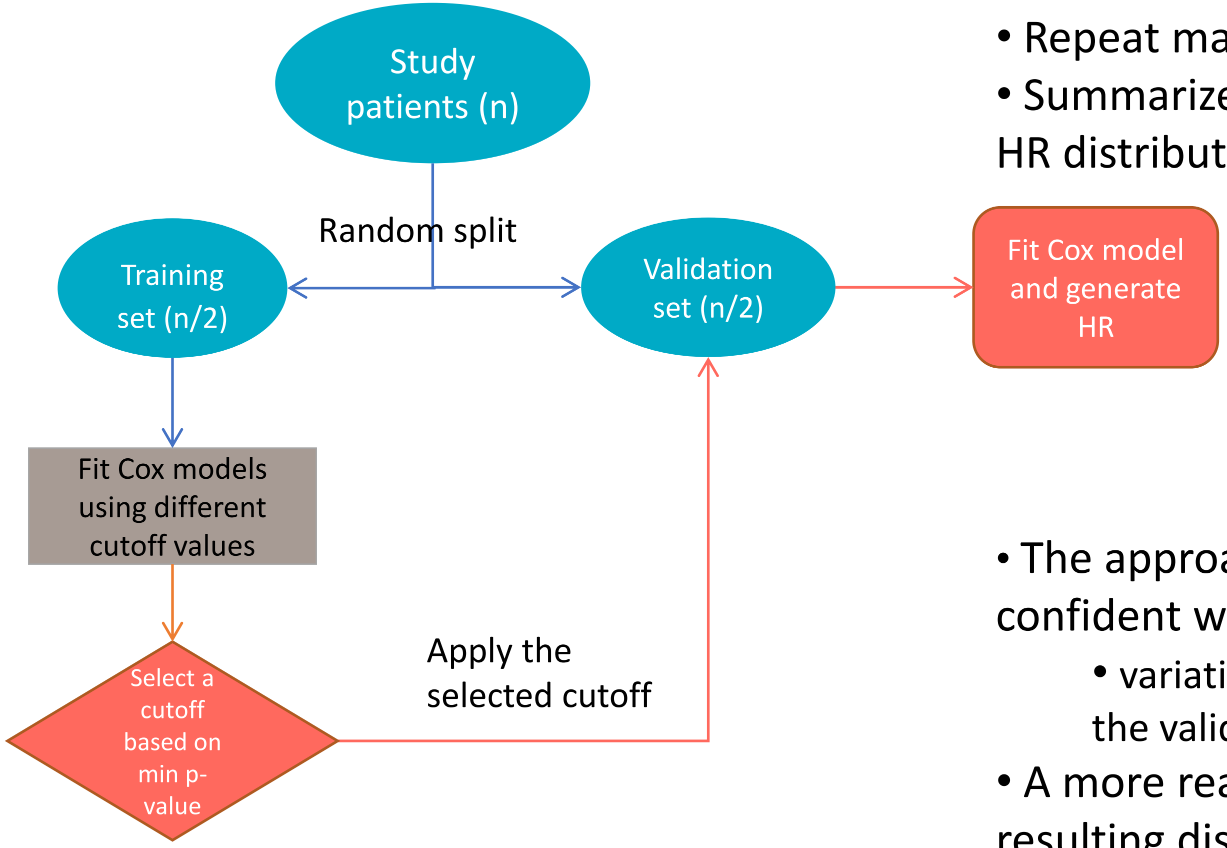
A new approach with multiple correlated factors*



- Permute the subgroup vector (preserves correlation) but not treatment, to create expected distribution of order statistics, to preserve the treatment effect in each bootstrap sample
  - Create sampling distribution of Z = (trt subgroup – trt overall)/sqrt [Var(trt subgroup)]
- Allows assessment of multiple correlated subgroups
  - Note, even if the age and gender distribution were independent, the female subgroup and the >50 age subgroup would be correlated. Correlated factors will lead to even higher correlations
- Examine whether observed extremes are outside expected distribution
  - Simulations confirm Type I Error and suggest greater power compared to interaction tests
  - In this example, the treatment effect in Tumour Grade 1 is larger than would be expected by chance alone allowing for the number of groups and their correlation

*Paper to be submitted by Aaron Dane, Amy Spencer, myself and David Svensson

Biomarker Cutoff Optimization – cross validation


Study patients (n)
Random split
Training set (n/2)
Validation set (n/2)
Fit Cox model and generate HR
Fit Cox models using different cutoff values
Select a cutoff based on min p-value
Apply the selected cutoff

- 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 of the extent of the resulting discrimination
  - Effect shrinkage in validation set

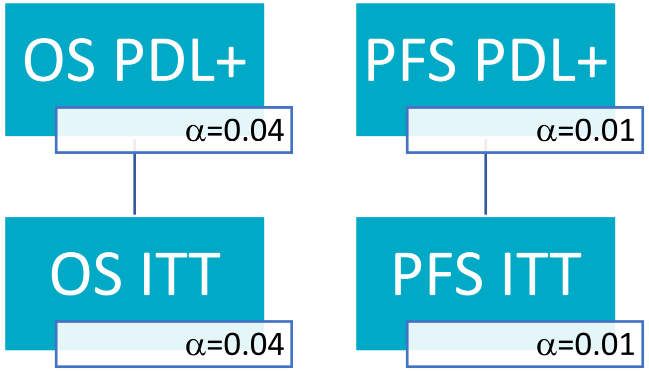
Confirming benefit

Normally uncertainty left

Importance of testing strategies in pivotal trials

A hypothetical example

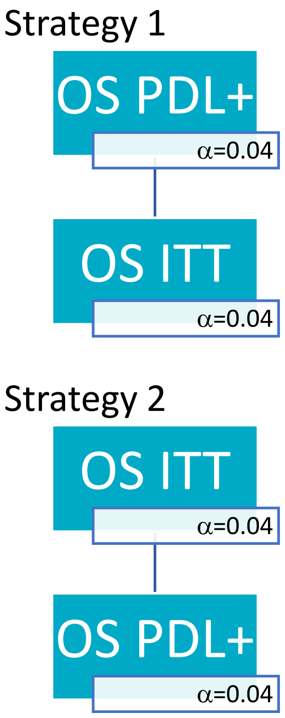
- Always critical that statisticians lead teams through different strategies
- Important all the team engage: too often seen as purely a stats issue when it can have a critical bearing on approval



- Could test PDL+ve population first as treatment effect likely to be bigger
- However,
  - The all-comers (ITT) population will have more events and may have more power
  - What if agent works as well in PDL-ve than expected?
    - Maybe control does badly in PDL-ve?
- Now will need to consider trade-offs
- Let’s assume PDL+ve group represents 30% of patients

Two different strategies for OS

same considerations for PFS



- PDL+ve
  - OS significant (p<0.04) if HR < 0.68 (80% power if HR=0.58)
  - 70% fewer events
- ITT
  - OS significant (p<0.04) if HR < 0.82

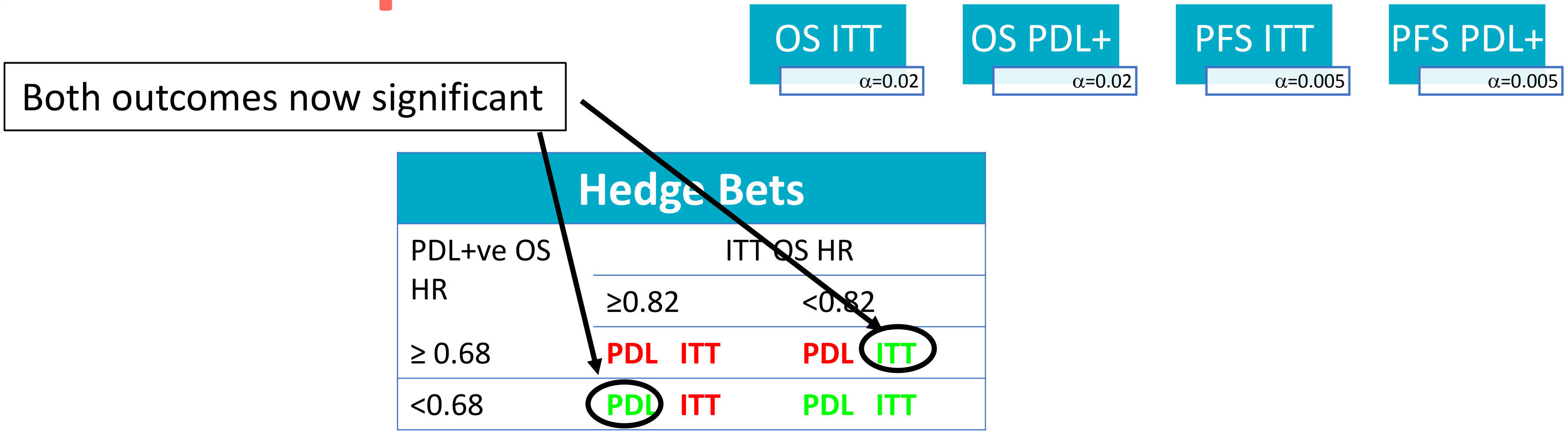
Trade?

Strategy 1		
PDL+ve OS HR	ITT OS HR	
	≥0.82	<0.82
≥ 0.68	PDL ITT	PDL ITT
<0.68	PDL ITT	PDL ITT

Strategy 2		
PDL+ve OS HR	ITT OS HR	
	≥0.82	<0.82
≥ 0.68	PDL ITT	PDL ITT
<0.68	PDL ITT	PDL ITT

Assume a true HR of 0.75 for OS with 80% power

Hedge bets again, and increase sample size to preserve power



OS ITT α=0.02

OS PDL+ α=0.02

PFS ITT α=0.005

PFS PDL+ α=0.005

Both outcomes now significant

Hedge Bets		
PDL+ve OS HR	ITT OS HR	
	≥0.82	<0.82
≥ 0.68	PDL ITT	PDL ITT
<0.68	PDL ITT	PDL ITT

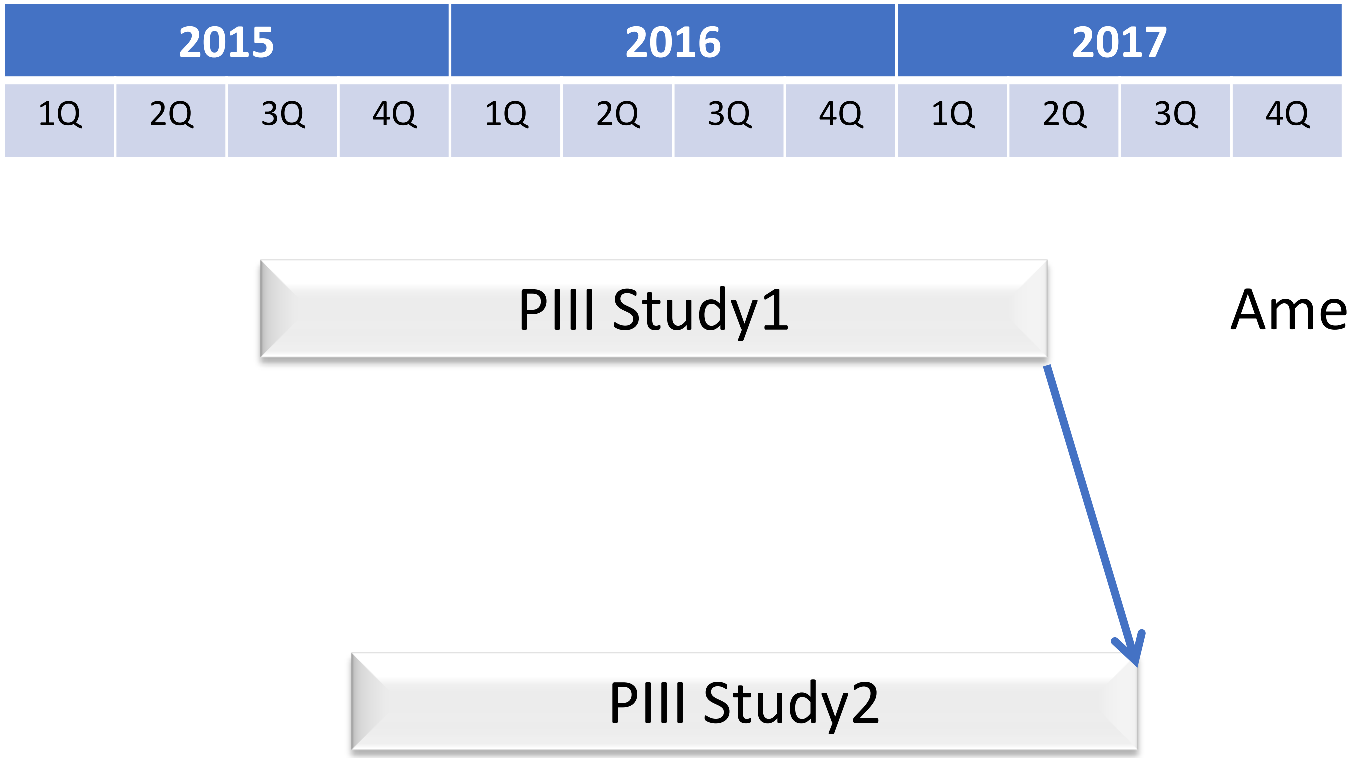
- By increasing the size of the trial to N=780, still have the same chance of demonstrating OS in ITT population as N=610 (80% power for single α=0.05 test of OS in ITT)
- So for an extra 27% of patients the trial has 4 ways to be positive instead of 1
    - And could lead to approval if any of those 4 were positive
    - Although if only ITT was positive, the indication may still be restricted if it was clear this result was driven by a more selective PDL population
- These results tend to surprise non-statistical colleagues who expect that sample size would need increasing by more

Enrichment

- Sometimes we consider enriching the trial population so that the % of patients recruited in a given subgroup is > than its prevalence in the population
  - Incidentally, this often seems to get confused with stratification with non-statistical colleagues
- This will increase power in the subgroup but:
  - the trial will take longer than it would
  - will mean, at some point, patients in the complement of subgroup will be prevented from being randomised
- Does lead to questions of how efficacy should be estimated in the full ‘ITT’ population, especially if efficacy depends on subgroup
  - An unweighted average may over-estimate efficacy
  - Whereas a weighted average, $pT^{+} + (1\text{-}p)T^{-}$, with variance = $p^{2}Var(T^{+}) + (1\text{-}p)^{2}Var(T^{-})$, may be more appropriate

One pivotal trial to define subgroup population for the 2nd trial

A simple modification but could have a huge bearing on success



2015				2016				2017			
1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q	1Q	2Q	3Q	4Q

Amend SAP and alpha spending in Study2 if evidence of predictive subgroup from Study 1

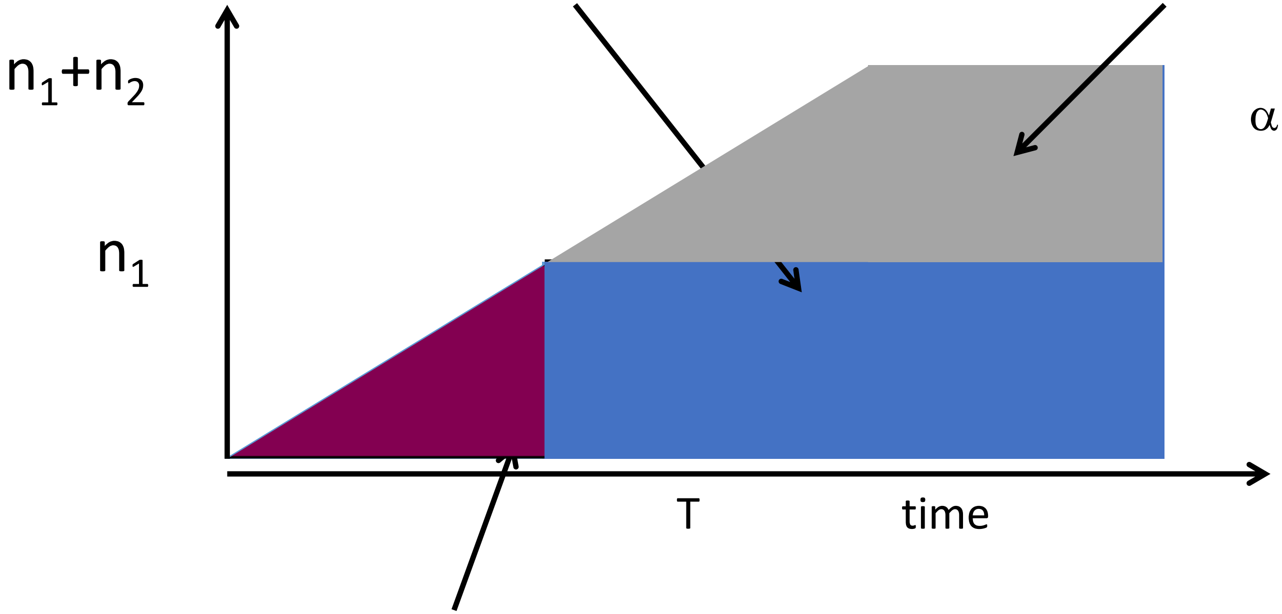
Learn and confirm?

Adaptive signature design

Learn and confirm within the same design

All patients and follow-up included in overall test at α_1

Subgroup hypotheses tested amongst n2 patients at α_2



$\alpha_1 + \alpha_2 = \alpha$

Subgroups explored amongst first n_1 patients followed to T

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

Adaptive sub-population design

Pre-defined hypothesis available

Flexibly define Stage 2 popn = overall, subgroup or both - or stop for futility

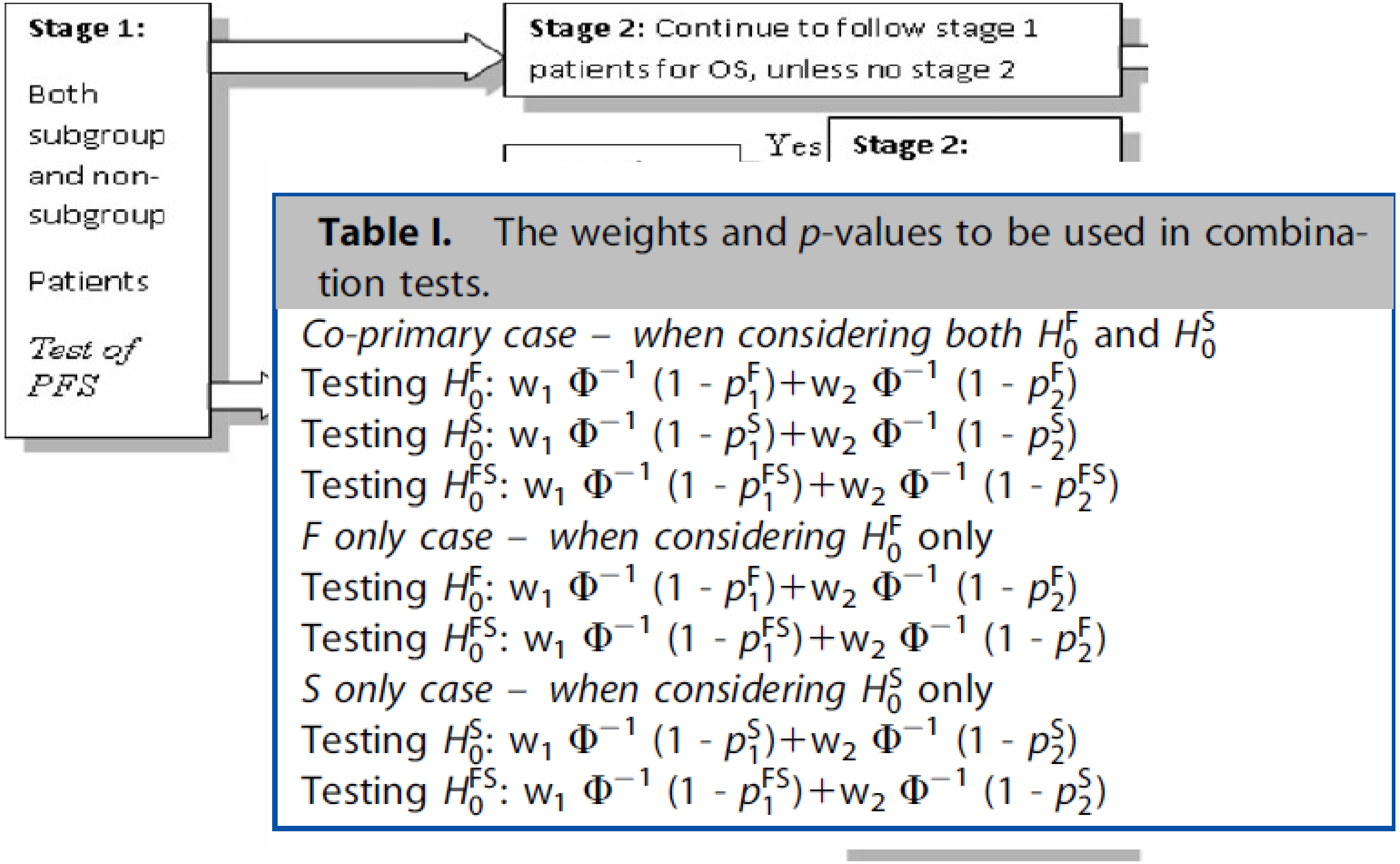

Stage 1:
Both subgroup and non-subgroup
Patients
Test of PFS
Stage 2: Continue to follow stage 1 patients for OS, unless no stage 2
Yes
Stage 2:

Table I. The weights and p-values to be used in combination tests.

Co-primary case – when considering both H_0^F and H_0^S

Testing H_0^F: $w_1\ \Phi^{-1}\ (1 - p_1^F) + w_2\ \Phi^{-1}\ (1 - p_2^F)$

Testing H_0^S: $w_1\ \Phi^{-1}\ (1 - p_1^S) + w_2\ \Phi^{-1}\ (1 - p_2^S)$

Testing H_0^{FS}: $w_1\ \Phi^{-1}\ (1 - p_1^{FS}) + w_2\ \Phi^{-1}\ (1 - p_2^{FS})$

F only case – when considering H_0^F only

Testing H_0^F: $w_1\ \Phi^{-1}\ (1 - p_1^F) + w_2\ \Phi^{-1}\ (1 - p_2^F)$

Testing H_0^{FS}: $w_1\ \Phi^{-1}\ (1 - p_1^{FS}) + w_2\ \Phi^{-1}\ (1 - p_2^F)$

S only case – when considering H_0^S only

Testing H_0^S: $w_1\ \Phi^{-1}\ (1 - p_1^S) + w_2\ \Phi^{-1}\ (1 - p_2^S)$

Testing H_0^{FS}: $w_1\ \Phi^{-1}\ (1 - p_1^{FS}) + w_2\ \Phi^{-1}\ (1 - p_2^S)$

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

Case study

Olaparib BRCA

An evolving phase II in ovarian cancer

Retrospectively seeking registration on the basis of a trial intended as phase II

- In 2008, a randomised phase II study was started comparing maintenance olaparib (capsules) vs placebo in 265 patients
  - Platinum sensitive ovarian cancer therapy
  - ≥ 2 prior platinum therapies
  - Response to most recent platinum therapy
  - These were expected to identify patients more likely to benefit
- Olaparib was predicted to be most efficacious in patients with a germline BRCA mutation that can be detected in the blood
  - In original analysis, status was ascertained in only ~35% of patients
- Overall, PFS HR (95% CI) = 0.35 (0.25, 0.49)*
  - Early data but no evidence of a survival improvement
  - HR point estimate between 0.1 and 0.2 in known BRCA+ve
- Key regulatory issues were typical of maintenance setting
  - No evidence of OS
  - These patients would normally receive a break from therapy and will now experience AEs
  - As well as methodological issues related to this study originally planned as a phase II study

*NEJM 2012 366 1382-92

Retrospectively establishing BRCA status

- Germline BRCA (gBRCA) status was subsequently retrospectively established in 80% of patients
  - A Further 16% of patients, BRCA status was established from tumour biopsies
- These data submitted for regulatory approval

	HR (95% CI)	
	gBRCA+ve	gBRCA WT*
PFS	0.18 (0.10, 0.31)	0.54 (0.34, 0.85)
OS	0.73 (0.45, 1.17)	0.99 (0.63, 1.55)

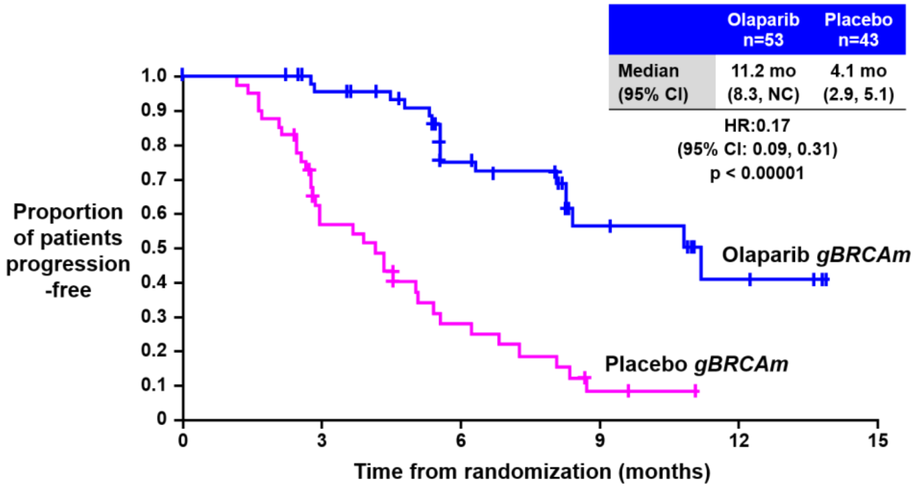
*WT = Wild-Type or gBRCA -ve
Ledermann Lancet oncology 2014

FDA Advisory Committee (ODAC) 2014

- Submission went to an ODAC in June 2014
- FDA commented
  - *there are uncertainties related to the validity and the reproducibility of the magnitude of effect seen in Study 19*
  - *In a pre-specified analysis of a retrospectively identified subgroup no alpha adjustments were made for multiplicity introduced by analyzing multiple endpoints (excluding overall survival), or analyses within the BRCA subgroups.*
- All of this is a fair criticism by FDA from a study originally planned as phase II
- Key factors was replication in another study
  - Closely related randomised PII: olaparib given in combination with chemo and then as maintenance

BRCA+ve results from key and supportive studies

Key Phase II gBRCA+ve

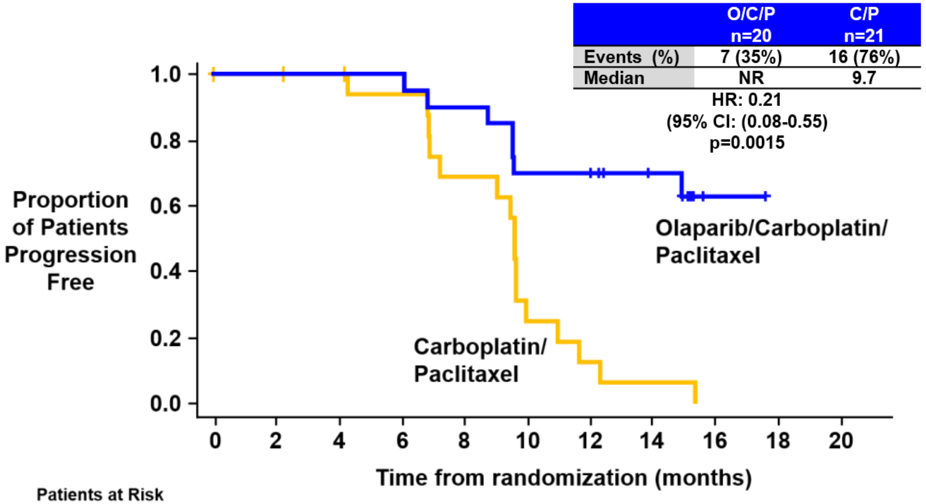


	Olaparib n=53	Placebo n=43
Median (95% CI)	11.2 mo (8.3, NC)	4.1 mo (2.9, 5.1)

HR:0.17
(95% CI: 0.09, 0.31)
p < 0.00001

Number at Risk						
Olaparib *gBRCAm*	53	44	26	11	4	0
Placebo *gBRCAm*	43	21	9	2	0	0

Supportive Study tBRCA+ve (tumour)

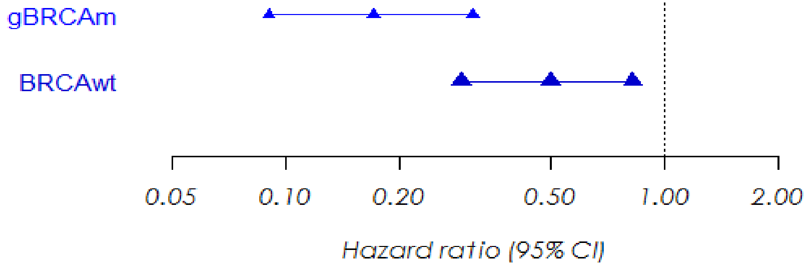


	O/C/P n=20	C/P n=21
Events (%)	7 (35%)	16 (76%)
Median	NR	9.7

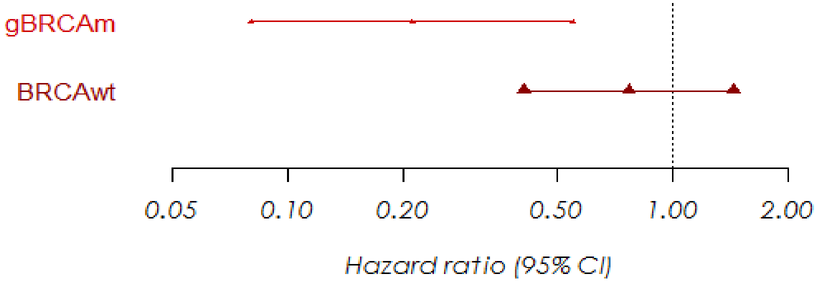
HR: 0.21
(95% CI: (0.08-0.55)
p=0.0015

Patients at Risk											
O/C/P	20	20	20	20	18	14	14	10	1	0	0
C/P	21	18	17	15	11	4	2	1	0	0	0

Interaction p=0.03



Interaction p=0.01



The Lancet Oncology , Volume 15 , Issue 8 , 852 – 861, 2014
The Lancet Oncology , Volume 16 , Issue 1 , 87 – 97, 2015
http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/OncologicDrugsAdvisoryCommittee/UCM402209.pdf

FDA also questioned whether the BRCA subgroup was randomized?

Could it be explained by an imbalance in prognostic factors?

- FDA: *The retrospective identification of the gBRCAm population did not appear to result in gross imbalances of known prognostic factors…. <u>but it is important to note that the loss of randomization </u>and the selection of a convenient sample of patients …. <u>may have led inadvertently to an unequal distribution of unknown factors that may have affected the study results</u>*
- Some confusion with lack of stratification
  - Which does not mean patients were not randomly assigned
  - Just in the same way as subgroup formed by age would be expected, on average, to be balanced for confounders
- There are legitimate concerns but not related to lack of randomization
  - gBRCA (blood sample) was not a primary endpoint – but see replication
  - This issue is more of a missing data problem
    - gBRCA status not established in 20% of patients
    - However, in a further 16% of patients BRCA status was established from tumour tissue, and tumour BRCA (tBRCA) status was a strong predictor for gBRCA – and the supportive study showed nearly the same effect in tBRCA +ve
- More generally there seems to be some confusion that only stratification can achieve balance
  - Just makes it more likely
  - Useful ref Kaiser Pharm Stats (2013 43-47)

Outcome

- Olaparib given accelerated approval in Dec 2014 but in a different indication in US
    - ~30% response rate seen in late line gBRCA patients
- Olaparib was approved in this population by EMA in Jan 2015
- Subsequently the phase II resulted in OS data of HR (95% CI) = 0.73 (0.55, 0.96) with further follow-up
    - Not statistically significant due to alpha spent earlier
- Phase II repeated in gBRCA only with a new tablet formulation
    - HR = 0·30 [95% CI 0·22-0·41] – 19.1m vs 5.5m medians
    - Full approval, including in the Wild-Type population

Pujade-Lauraine Lancet Oncol 2017 1274-84

Post-treatment biomarkers – an aside

- We haven't discussed the possibility of using a biomarker that's measured in response to therapy as a predictive biomarker
  - Eg Rash and EGFR therapies, response to a PET scan

- This is a very complex issue
  - Analyses by such a biomarker (rash) can easily mislead
    - The biomarker may just be a good surrogate for initial prognosis
  - Designs may be possible* though complex
    - In progressive diseases, need to establish that delay in switching to more appropriate therapy does not result in long term harm to the patient
    - Appropriate comparisons, so that differences can be attributed to treatment and not other unmeasured confounders

* Stone *Pharmaceutical Statistics* 2014 13: 214-221

Conclusions

Development of agents based on predictive biomarkers becoming increasingly important

- Almost the default position in most of oncology

Many important factors

- Design
- Identification
- Interpretation