Multiplicity

Is it of value to make it so complicated?

Andrew Stone
Oncology TA Statistical Expert, AstraZeneca
EMA workshop on multiplicity – Nov 2012

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.

Your car needs some more brake fluid

Would you like Strong or Weak control Sir?

Is it of value to make multiplicity so complicated?

- **Increased complexity in multiplicity procedures, to satisfy ‘strong control’, is in danger of becoming self-defeating**
  - So opaque that it is disregarded in interpretation

- **We should question whether the extra complexity is worth it**
  - Does strong control add value to our assessment of medicines?

- **Should we take a more considered approach in order to:**
  - **Decide whether a drug should be licensed?**
  - **Provide analyses that inform prescribers as to the nature of the benefit and risks?**

Stepping Back

Why do we do statistical analyses?

What are the key roles that our statistical analyses play?

1. ***Decide whether a drug should be licensed***
2. ***If yes, provide analyses that inform prescribers the nature of the benefit and risks of that agent***

Therefore our analysis approach should be congruent with those aims

- but also mindful that they are not unnecessarily complicated and as a result hinder interpretation

Whether a drug should be licensed
Key question: was the trial positive?

Consider:

- Whether the results on the primary endpoints are inconsistent with the play of chance
  - Statistical convention*, per trial, for a trial to be +ve there must be a < 2.5% chance of a false +ve
- Whether the design, conduct or analysis of the trial is biased in a way that may have inflated the false probability despite what the p-value indicates

*Requiring 2 +ve PIII means there is 0.025^2 =0.000625 false +ve, or approval on a single trial might use a lower alpha than 0.025

Simple examples exerting only Weak Control[##]

Even though the probability of falsely claiming a positive trial is controlled

- A +ve trial if ***either*** of 2 analyses are statistically significant
  - For example two experimental arms, use a significance level of 1.25% (1-sided)[#] per comparison*

- A +ve trial if ***both*** of 2 analyses are statistically significant
  - For example two co-primary endpoints, use a significance level of 2.5% (1-sided) per endpoint

- Without further measures, this however only exerts so called ‘Weak control’ of Type I error
  - Even though the probability of falsely claiming a positive trial is controlled

* Allowing for the known correlation between comparisons, slightly higher significance levels could be applied and still control Type I Error between the 2 endpoints

[#] unless stated significance levels are 1-sided throughout the presentation

[##] assuming all secondary endpoints were also tested at 2.5% 1-sided

So what's Strong Control?

- 'The probability of rejecting any (i.e. one or more) true null[1] H[2] is at most 2.5%, irrespective of how many and which Hs that actually are true or false.'

- In other words, amongst those claims, where in truth there is no treatment effect[3] (incidentally, we'll never know which those are!), there is a <2.5% chance of falsely claiming an effect on at least one of them

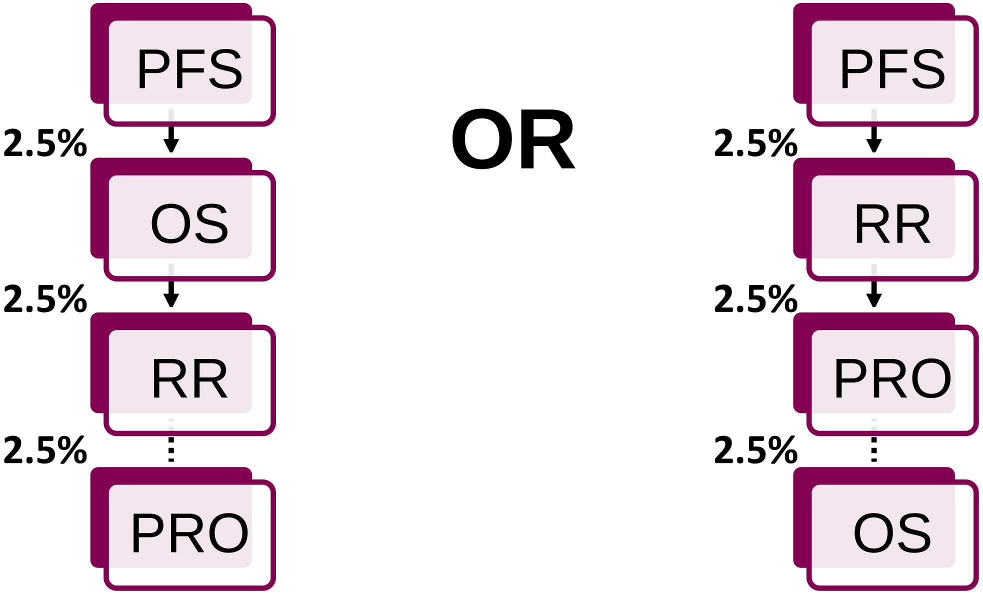
- Therefore, only requires adjustment amongst secondary endpoints if it turns out there truly is an effect on the primary endpoint*
  - Which seems inconsistent with the necessary equipoise to be prepared to randomise

[1] True null is one where there is truly no treatment effect (in a superiority trial at least)
[2] H: hypothesis
[3] in a superiority trial
* If the null H was true for the primary endpoint, then even if <u>all</u> secondary endpoints were tested at 2.5% and they were <u>only</u> tested if the primary endpoint was significant, the probability of falsely rejecting one of the secondaries is < 2.5% - as they are only tested at most 2.5% of the time; when the primary is significant

Strong control in practice

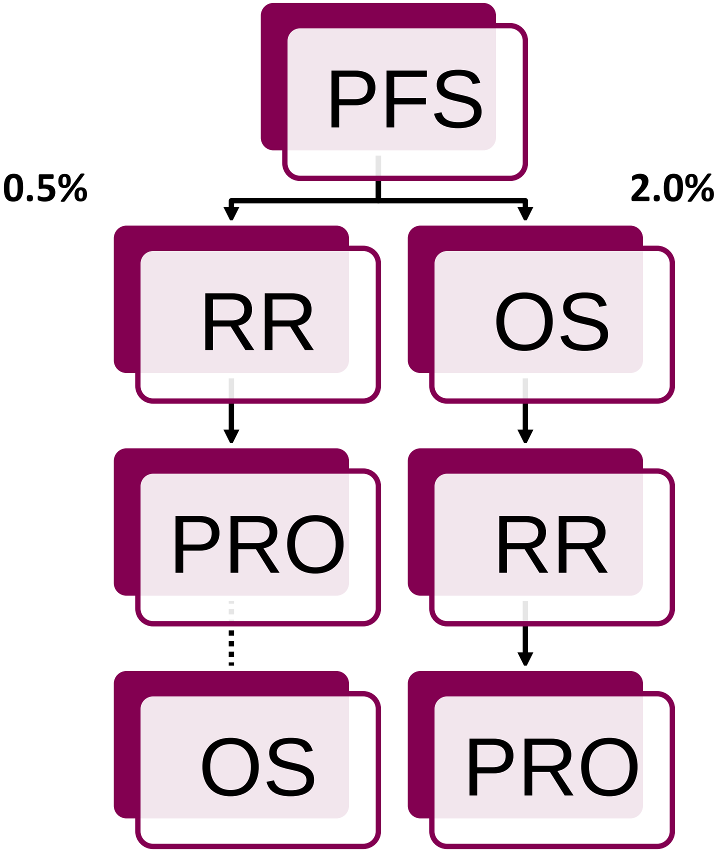
- Boundaries between primary and secondary endpoints blurred
- Requires ranking of endpoints and placing of ‘bets’ on:
  - The endpoints that are most likely to be significant

  OR
  - The most important endpoints even if they are less likely to be significant



PFS = disease progression (primary endpoint),
OS = survival, RR = response rate, PRO = symptoms or QOL

So I’ll spread my bets



Complexity soon escalates

The temptation though is that it is fascinating to identify the perfect schema

Example: Two experimental arms

3 arms: **Control**
Exp.+ Control (*Combo*)
Exp. (*Mono*)

- If $p_{combo} < 1.25\%$ do we:

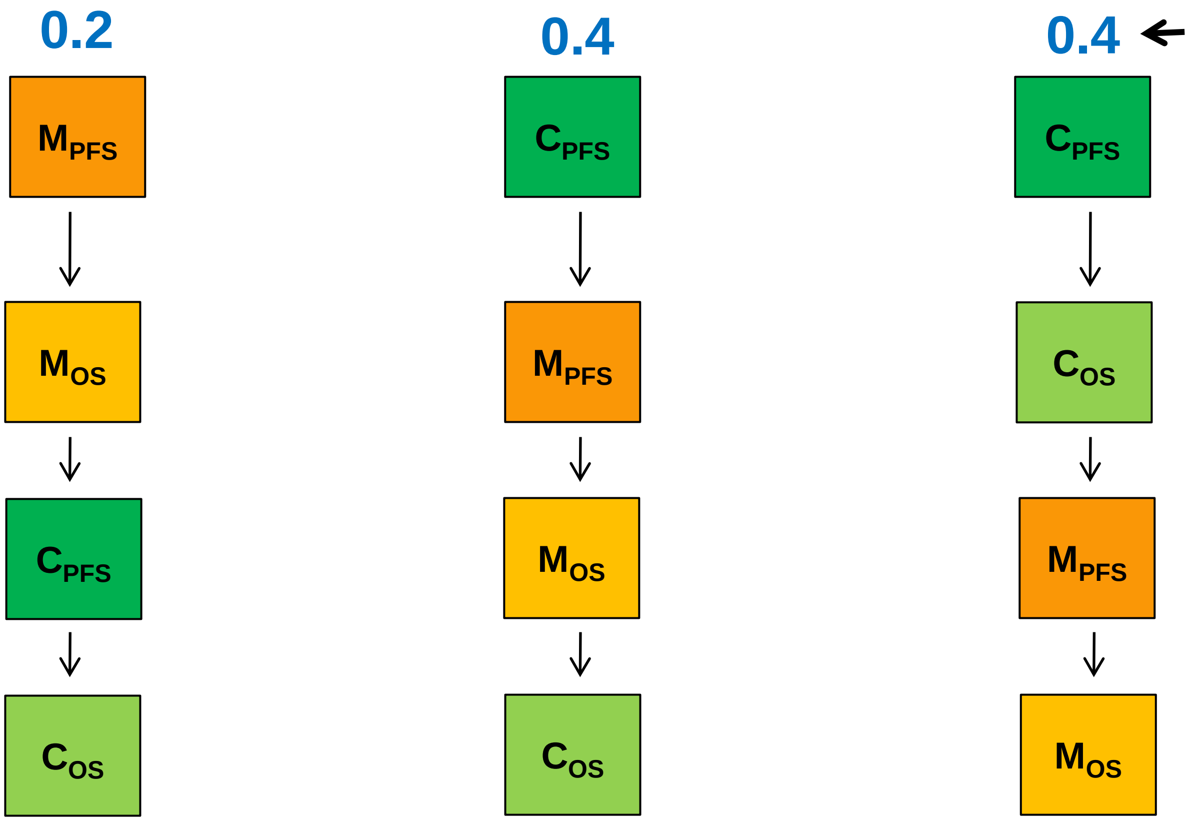
1. Pass alpha to mono and test mono at 2.5% (Holm procedure amongst primary endpoints)?
2. Pass some proportion of alpha instead to a combo secondary endpoint?

Otherwise if $p_{mono} > 5\%$ then can we make claims on any combo secondary endpoints even if $p<0.00001$??

Brief example

- Various methodologies; intuitively most appealing
  - Burman et al: Statistics in Medicine 2009: 28: 739-761
  - Bretz et al: Statistics in Medicine 2009: 28: 586-604
- Re-cycling (Burman et al) methodology highlighted here
- For each comparison (mono and combo)
  - 2 endpoints considered OS and PFS

Testing set-up and p-values



0.2

M_{PFS} → M_{OS} → C_{PFS} → C_{OS}

0.4

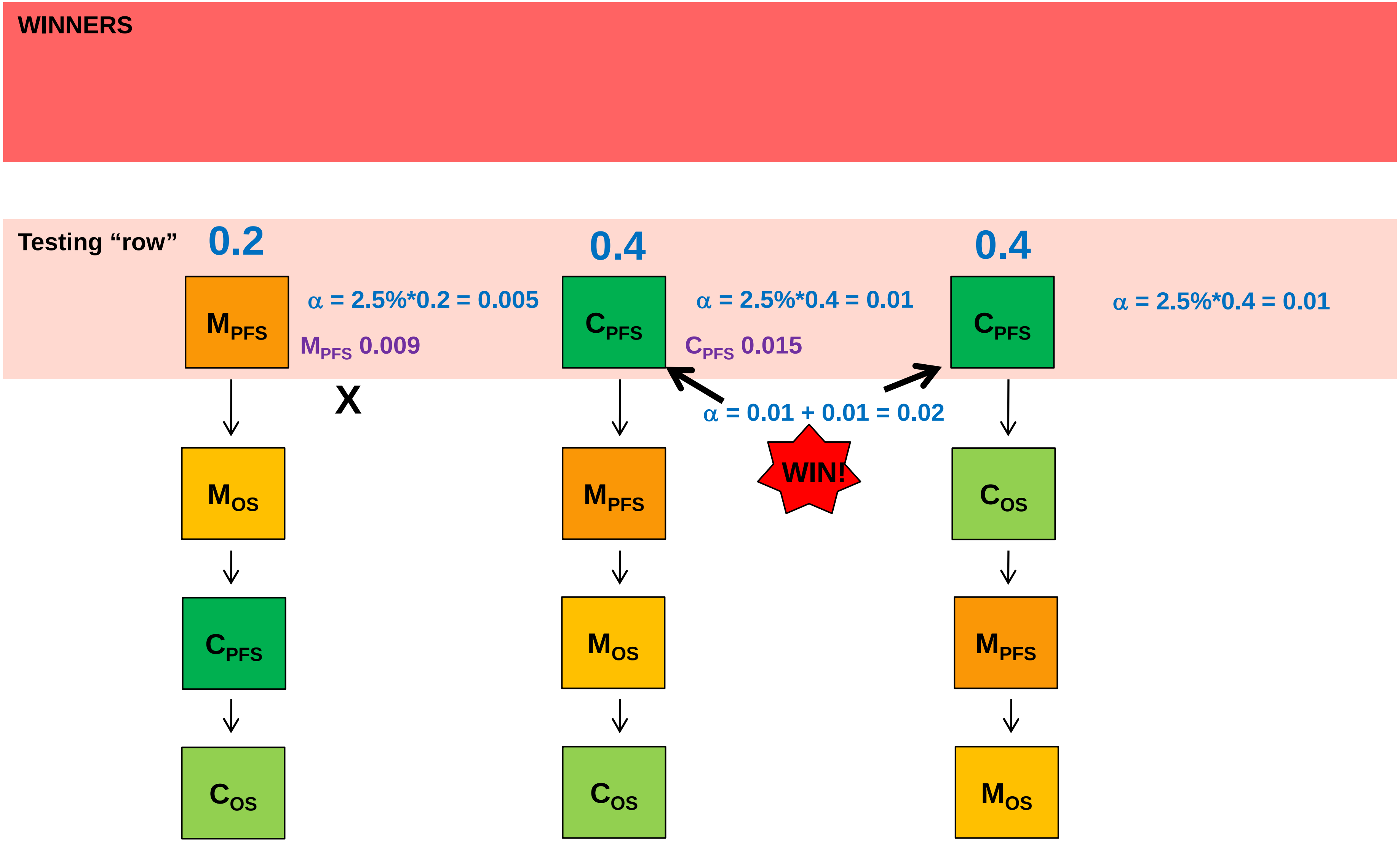
C_{PFS} → M_{PFS} → M_{OS} → C_{OS}

0.4

C_{PFS} → C_{OS} → M_{PFS} → M_{OS}

Test mass: proportion of alpha available

1-sided p-values: M_{PFS}=0.009, M_{OS}=0.1, C_{PFS}=0.015, C_{OS}=0.02



WINNERS

Testing “row”

0.2

M_{PFS}

$\alpha = 2.5\%*0.2 = 0.005$

M_{PFS} 0.009

X

M_{OS}

C_{PFS}

C_{OS}

0.4

C_{PFS}

$\alpha = 2.5\%*0.4 = 0.01$

C_{PFS} 0.015

$\alpha = 0.01 + 0.01 = 0.02$

WIN!

M_{PFS}

M_{OS}

C_{OS}

0.4

C_{PFS}

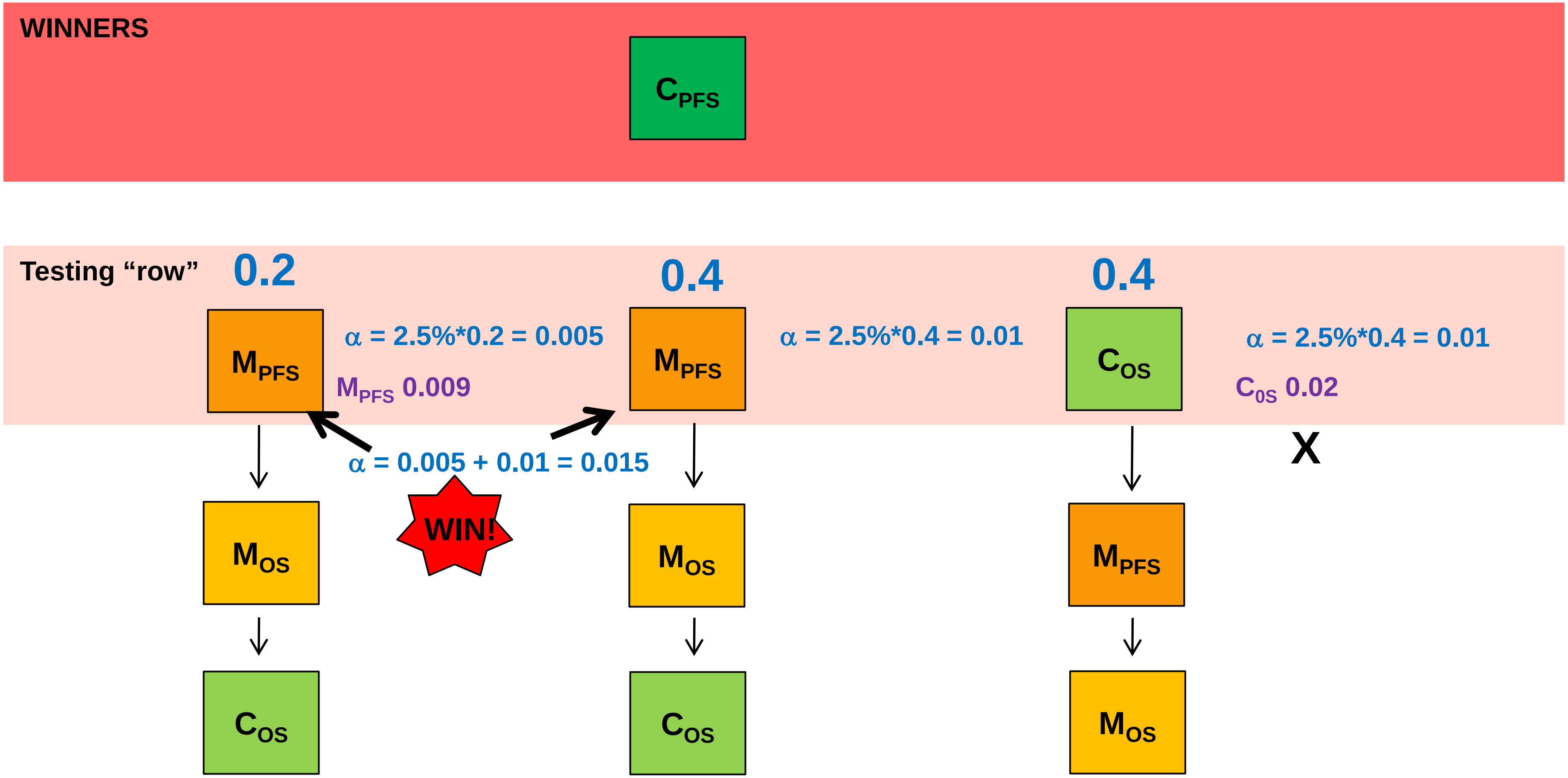
$\alpha = 2.5\%*0.4 = 0.01$

C_{OS}

M_{PFS}

M_{OS}

1-sided p-values: M_{PFS}=0.009, M_{OS}=0.1, C_{PFS}=0.015, C_{OS}=0.02



WINNERS

C_{PFS}

Testing “row”

0.2

M_{PFS}

α = 2.5%*0.2 = 0.005

M_{PFS} 0.009

M_{OS}

C_{OS}

α = 0.005 + 0.01 = 0.015

WIN!

0.4

M_{PFS}

α = 2.5%*0.4 = 0.01

M_{OS}

C_{OS}

0.4

C_{OS}

α = 2.5%*0.4 = 0.01

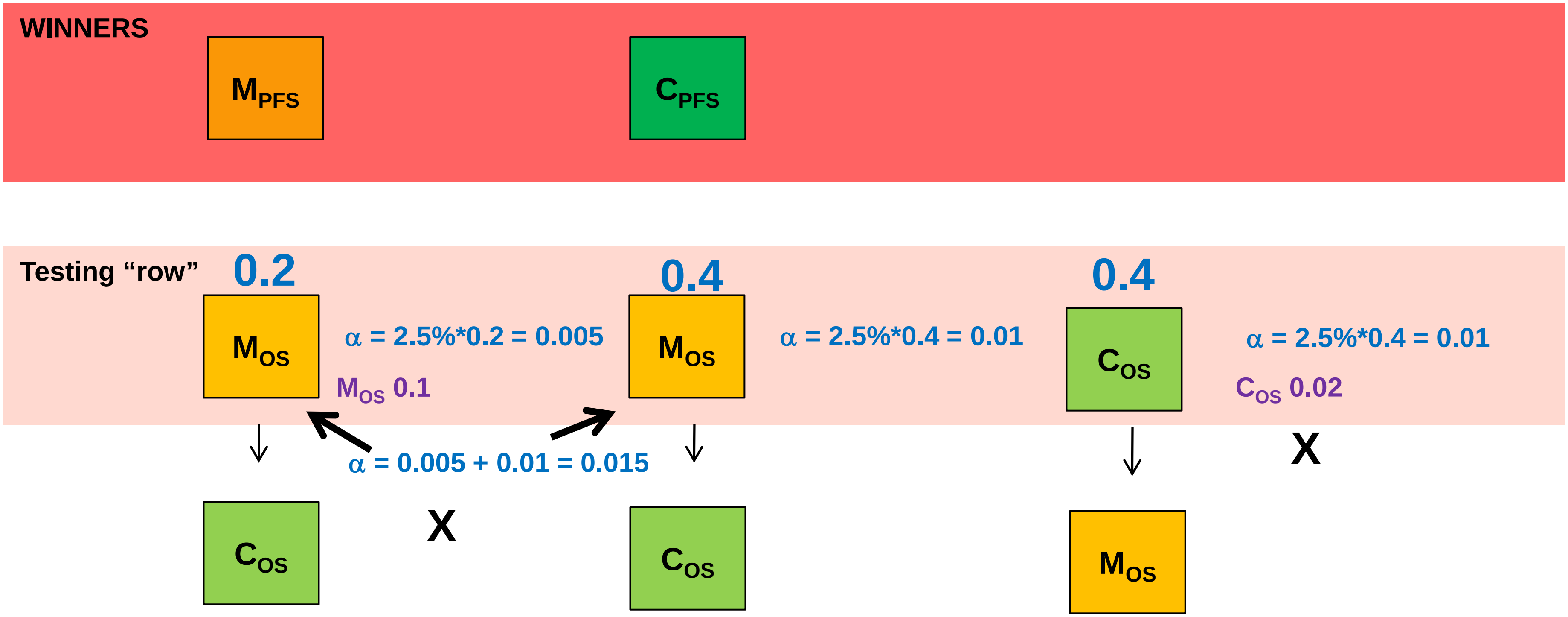
C_{0S} 0.02

X

M_{PFS}

M_{OS}

1-sided p-values: M_{PFS}=0.009, M_{OS}=0.1, C_{PFS}=0.015, C_{OS}=0.02



WINNERS

M_{PFS}

C_{PFS}

Testing “row”

0.2

M_{OS}

$\alpha = 2.5\%*0.2 = 0.005$

M_{OS} 0.1

0.4

M_{OS}

$\alpha = 2.5\%*0.4 = 0.01$

0.4

C_{OS}

$\alpha = 2.5\%*0.4 = 0.01$

C_{OS} 0.02

$\alpha = 0.005 + 0.01 = 0.015$

X

X

C_{OS}

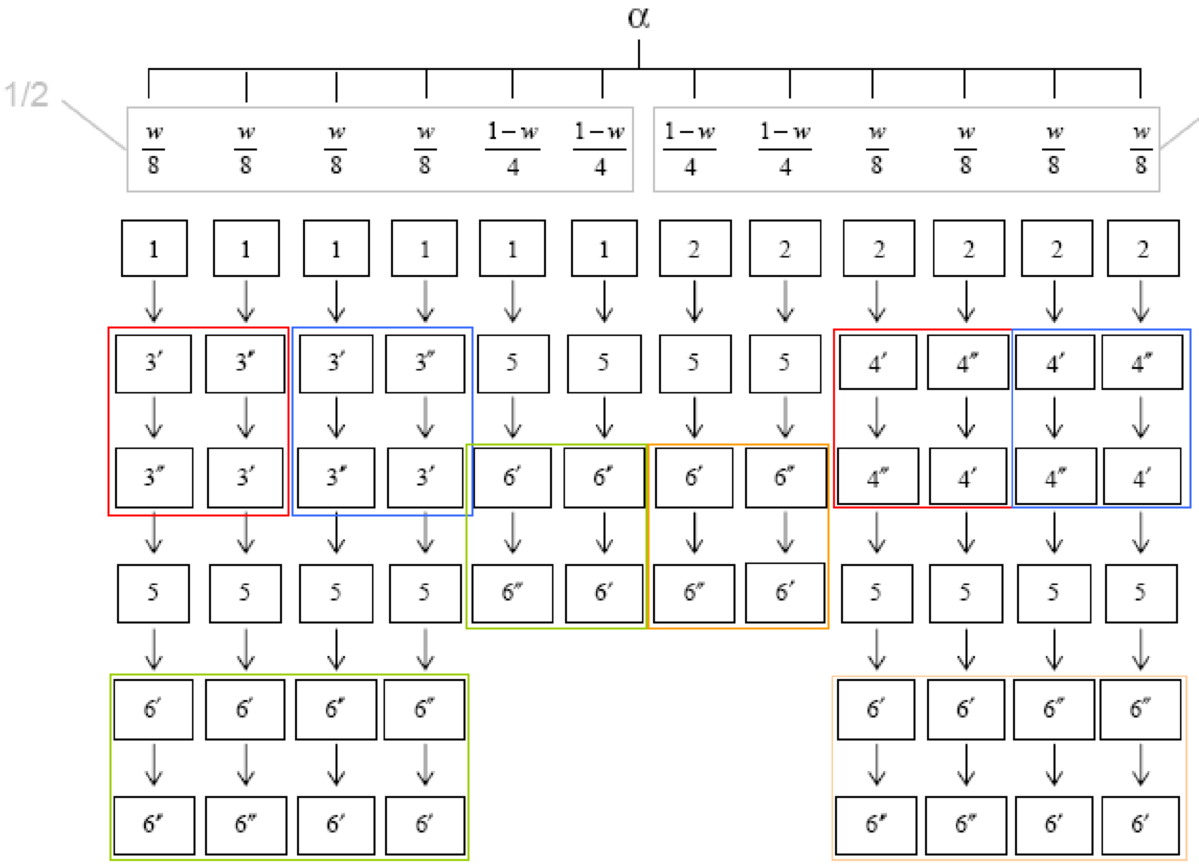
C_{OS}

M_{OS}

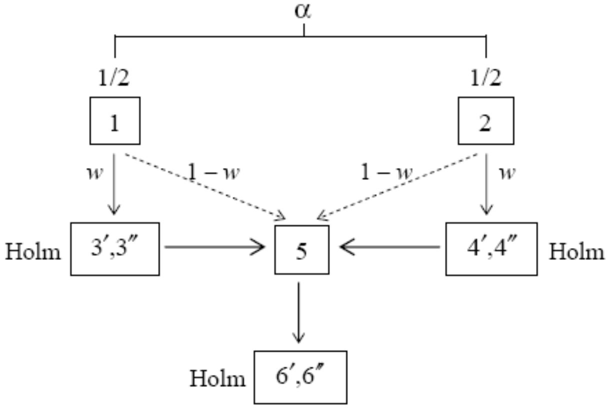
1-sided p-values: M_{PFS}=0.009, M_{OS}=0.1, C_{PFS}=0.015, C_{OS}=0.02

Yikes!

Now I have 3 experimental dose arms, a primary and 2 secondary endpoints per comparison


α
1/2
$\frac{w}{8}$ $\frac{w}{8}$ $\frac{w}{8}$ $\frac{w}{8}$ $\frac{1-w}{4}$ $\frac{1-w}{4}$ $\frac{1-w}{4}$ $\frac{1-w}{4}$ $\frac{w}{8}$ $\frac{w}{8}$ $\frac{w}{8}$ $\frac{w}{8}$
1 1 1 1 1 1 2 2 2 2 2 2
3′ 3″ 3′ 3″ 5 5 5 5 4′ 4″ 4′ 4″
3″ 3′ 3″ 3′ 6′ 6″ 6′ 6″ 4″ 4′ 4″ 4′
5 5 5 5 6″ 6′ 6″ 6′ 5 5 5 5
6′ 6′ 6″ 6″ 6′ 6′ 6″ 6″
6″ 6″ 6′ 6′ 6″ 6″ 6′ 6′

Or more simply, but you still have to work through the full diagram in practise

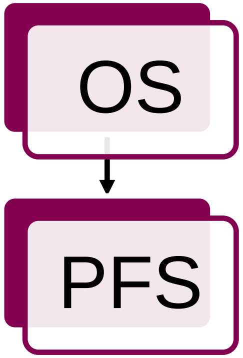


1 = dose1 primary; 2 = dose2 primary; 3’,3’’ = dose1 secondaries

4’,4’’ = dose2 secondaries, 5 = dose3 primary, 6’,6’’ = dose3 secondaries

Let’s up the ante

Group Sequential Designs (GSDs)



- Primary endpoint tested twice and alpha controlled (O’Brien/Fleming or Pocock)

- If primary endpoint is significant at the interim (or final) analysis can the secondary be tested at 2.5% (1-sided)?
  - Regardless of when the primary endpoint is rejected

- No: Hung[1] highlighted under strong control there *might* be inflation
  - For some combinations of true treatment effects and strong correlations between endpoints

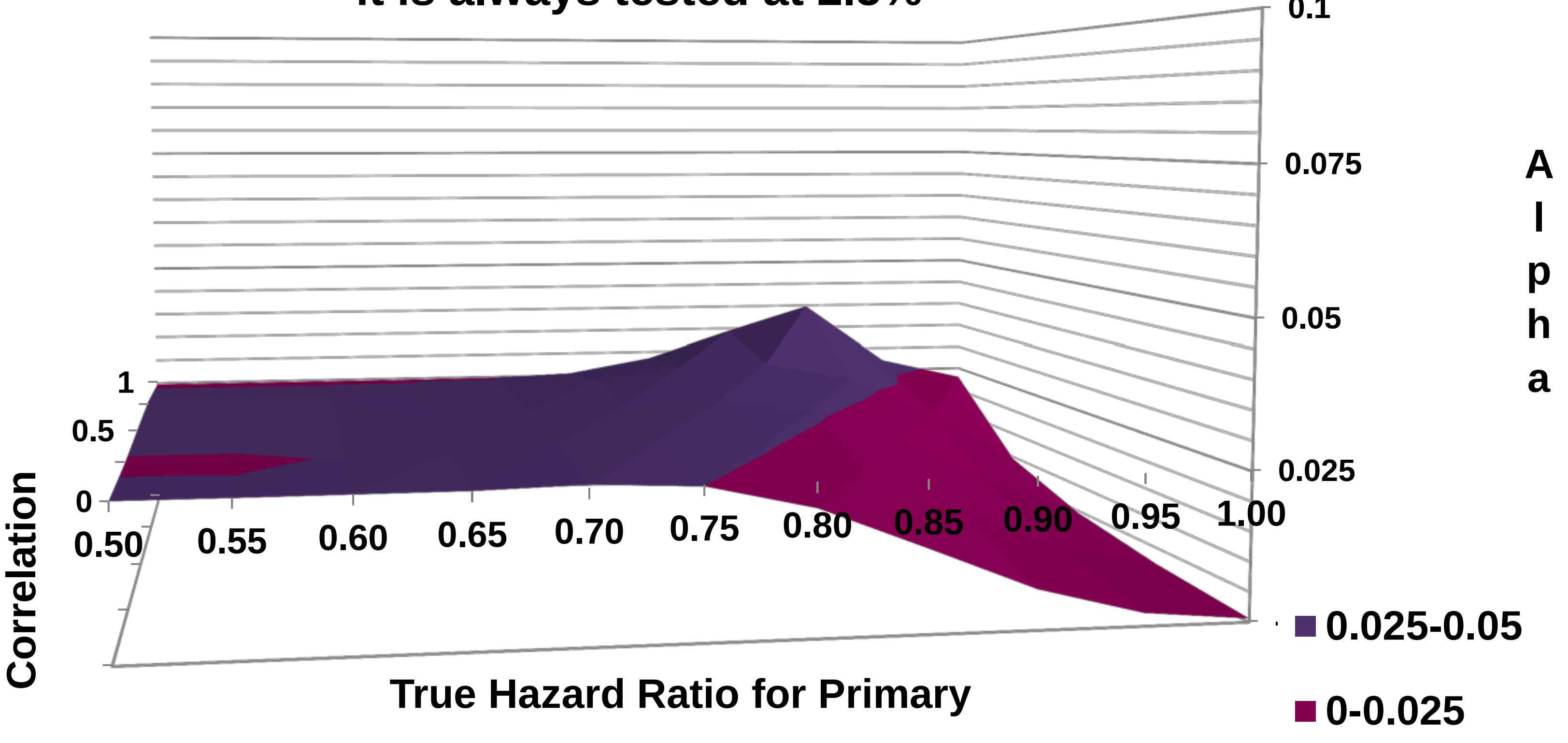
- Glimm[2] & Tamhane[3] provide a framework and a solution

- ***To achieve strong control, need to consider all possible permutations of the true treatment effect and correlation***

[1] Hung et al, J Biopharm Stat 2007: 17: 1201-1210
[2] Glimm et al, Stat Med 2009: 29:219-228
[3] Tamhane et al, Biometrics 2010: 1174-1184

Minimal inflation of alpha in GSDs in all but very specific and unlikely situations

Probability of falsely rejecting the secondary endpoint if it is always tested at 2.5%



Uses LanDemets 1-sided $\alpha^1(t) = 2\text{-}2F(z_{0.5\alpha}/\sqrt{t})$ - approximates O'Brien/Fleming
t = proportion of information for interim, normal deviates: primary interim,=-2.9625, final=-1.9685, secondary interim & final=-1.96, analyses at 350 and 700 events, max type I error occurs when true HR=0.9
Assume same proportion of information at interim for both PFS and OS
Based on 50000 values simulated per correlation HR combination

So how serious is inflation in practise?

- More than negligible inflation of alpha only occurs with correlations between primary and secondary close to 1
  - And only then for true treatment effects on the primary within a tight range

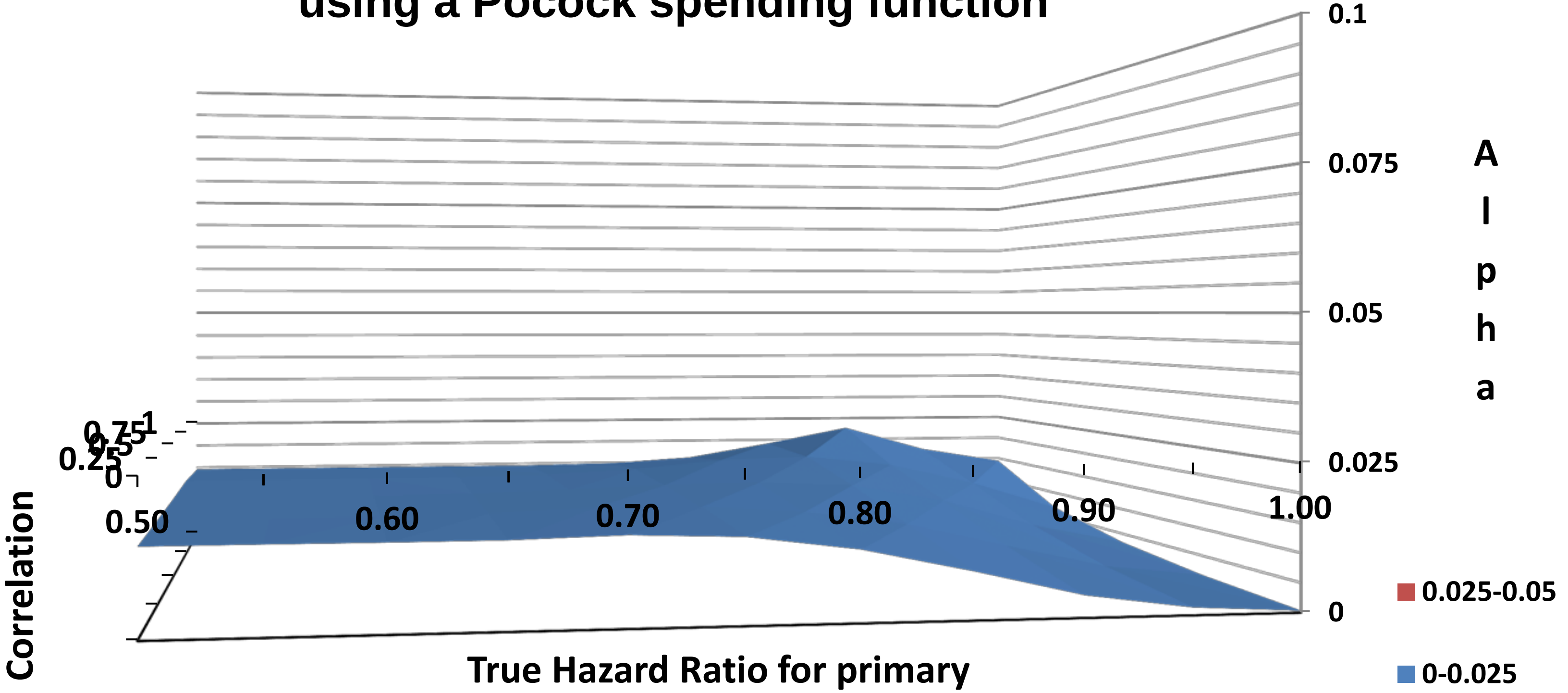
- Correlation between test statistics for OS and PFS even when temporally close
  - 0.5 3rd line NSCLC*, n=900 (median PFS=2m, median OS=6m)
  - 0.54 2nd line NSCLC, n=1300
  - 0.54 2nd line colorectal cancer, n=200

- With a correlation ≤ 0.5 the maximum alpha is 2.8% when the HR=0.8#.

AZ data: * non-small cell lung cancer, #amongst HRs of 1 to 0.5 by 0.05

A possible solution is to define a spending function for the secondary endpoint too

Probability of falsely rejecting the secondary endpoint if if its alpha, 2.5%, is split between the interim and primary using a Pocock spending function



Uses LanDemets 1-sided $\alpha^1(t) = 2-2F(z_{0.5\alpha}/\sqrt{t})$ for primary endpoint approximates O'Brien/Fleming, Pocock for secondary α = 0.0294 for both analyses, t = proportion of information for interim

Normal deviates: primary interim =-2.9625, primary final =-1.9685, secondary interim and final =-2.1781, analyses at 350 and 700 events, max type I error occurs when true HR=0.9

Assume same proportion of information at interim for both PFS and OS

Based on 50000 values simulated per correlation HR combination

Oh and…

- What if there are multiple dose or experimental arms and interim analyses?
  - The complexity level is ramped up even more

- What if the primary endpoint, OS, is analysed more than once but the secondary endpoints, PFS & RR, are collected once:
  - Do I then need to adjust them even with strong control?
  - And can they be tested twice, in case the final, but not interim, OS is significant?

Indulgence Over

Is all of this necessary?

As long as we:

- Rigorously and fully control Type I error amongst primary endpoints
- Don’t allow significant secondary endpoints to rescue a failed trial

Is it really necessary:

- That we need to control the probability of making false claims amongst those endpoints where the treatment truly has no effect – we’ll never know which those are!
  - Given the trial was positive, would now, in all likelihood, only be a subset of the secondary endpoints

Could we have created a framework that due to its complexity:

- Is hard for non-statisticians to understand the need for
- Becomes self-defeating
- And is largely disregarded when we come to:
  - Decide whether to license
  - And describe the nature of the benefit and risk

Brilique SPC presentation

All cause mortality below a non-significant endpoint in the hierarchy

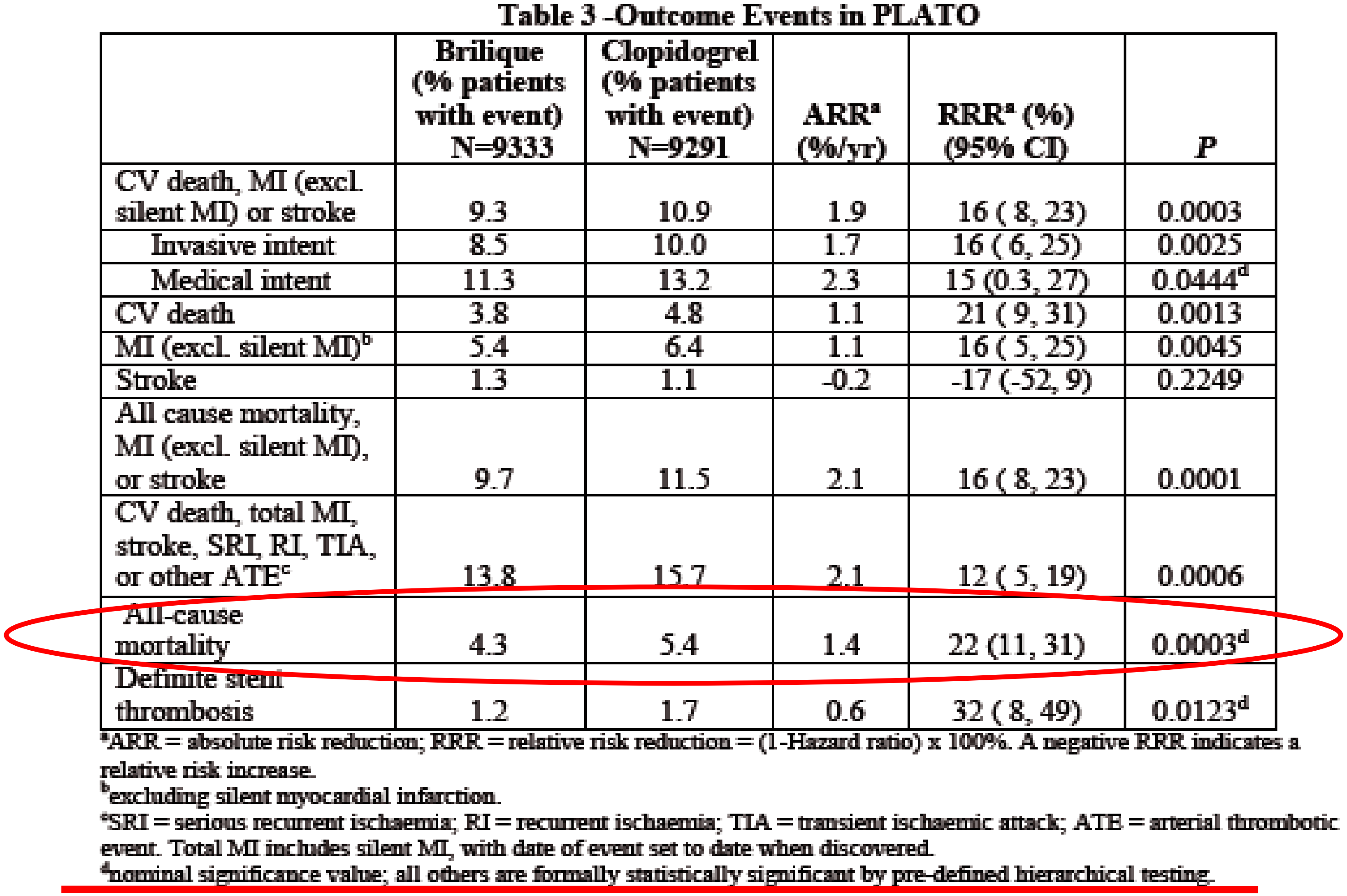


Table 3 -Outcome Events in PLATO

	Brilique (% patients with event) N=9333	Clopidogrel (% patients with event) N=9291	ARR[a] (%/yr)	RRR[a] (%) (95% CI)	*P*
CV death, MI (excl. silent MI) or stroke	9.3	10.9	1.9	16 (8, 23)	0.0003
Invasive intent	8.5	10.0	1.7	16 (6, 25)	0.0025
Medical intent	11.3	13.2	2.3	15 (0.3, 27)	0.0444[d]
CV death	3.8	4.8	1.1	21 (9, 31)	0.0013
MI (excl. silent MI)[b]	5.4	6.4	1.1	16 (5, 25)	0.0045
Stroke	1.3	1.1	-0.2	-17 (-52, 9)	0.2249
All cause mortality, MI (excl. silent MI), or stroke	9.7	11.5	2.1	16 (8, 23)	0.0001
CV death, total MI, stroke, SRI, RI, TIA, or other ATE[c]	13.8	15.7	2.1	12 (5, 19)	0.0006
All-cause mortality	4.3	5.4	1.4	22 (11, 31)	0.0003[d]
Definite stent thrombosis	1.2	1.7	0.6	32 (8, 49)	0.0123[d]

[a] ARR = absolute risk reduction; RRR = relative risk reduction = (1-Hazard ratio) x 100%. A negative RRR indicates a relative risk increase.

[b] excluding silent myocardial infarction.

[c] SRI = serious recurrent ischaemia; RI = recurrent ischaemia; TIA = transient ischaemic attack; ATE = arterial thrombotic event. Total MI includes silent MI, with date of event set to date when discovered.

[d] nominal significance value; all others are formally statistically significant by pre-defined hierarchical testing.

The associated EPAR

Appropriate interpretation of the data

Even though all cause mortality was not significant according to the hierarchical testing procedure:

‘All cause mortality was also significantly reduced (HR 0.78 (95% CI 0.69, 0.89); p=0.0003). ‘

‘The most important secondary endpoints are supportive for the primary endpoint, including all cause mortality’

OK- so what’s the alternative?

- Rigorously and fully controlled Type I Error amongst the primary endpoints
- Concentrate mostly on design, conduct and analysis measures to minimise possible bias so that Type I Error is actually controlled
- If the trial is positive, view the role of secondary endpoints as describing the nature of any benefit
- But be sensible about the analysis plan for secondary endpoints:
  - Group them according to the separate clinical questions (see next slide)
  - Within those, exercise alpha control via:
    - Nomination of a key secondary which acts as a gate-keeper for a fuller description (eg multiple timepoints, or multiple aspects of QOL)
    - Alpha control amongst related endpoints

Oncology example

Benefit?

Did the patient live longer?

OS – control alpha for repeated analyses

Nature of the benefit?

Did the underlying disease progress later?

Test Progression Free Survival (PFS) at 2.5%

Did the patient experience more tumour shrinkage?

Control alpha 2.5% amongst Response Rate and Duration of Response

Did the patient feel or function better?

Control alpha 2.5% amongst a suite of symptomatic and HRQOL endpoints

In conclusion

Whilst the whole multiplicity area and literature is intellectually fascinating:

- Let’s not lose sight of what we’re trying to achieve
- Question whether the complexity to achieve strict Strong Control of type I error is worth it, and indeed could become self-defeating

Maintain our focus on:

- Minimisation of bias
- With rigorous control of alpha amongst primary endpoints

Then take a more considered approach to the role that multiplicity takes in:

- Licensing of medicines
- And description of their benefits and risks

Confidentiality Notice